息没有关系。换句话说，正交对比指一组系数中的数字（如 C_{a1}，C_{a2}，C_{a3}，…，C_{ap}）与另一组系数中的数字（如 C_{b1}，C_{b2}，C_{b3}，…，C_{bp}）是完全无关的。

（一）正交对比的系数

我们可以将感兴趣的问题或对比用系数的方式来表示。例如，如果对“语音教学法（A1）与传统教学法（A3）是否有差别，语素教学法（A2）与传统教学法（A3）是否有差别，以及实验教学法与传统教学法是否有差别”这样一组对比感兴趣，可以用系数表示如下：

C_1	C_2	C_3
1	0	−1
0	1	−1
$\frac{1}{2}$	$\frac{1}{2}$	−1

然而，如果对另外一组对比感兴趣，例如“实验教学法与传统教学法是否有差别，语音教学法与语素教学法是否有差别”，则是使用另外一组系数表示的：

C_1	C_2	C_3
$\frac{1}{2}$	$\frac{1}{2}$	−1
1	−1	0

当我们对同一组数据提问两个或两个以上问题时，有时我们希望知道这两个问题之间是否独立。评价对比是否独立的问题就是对比的正交性问题。判断两个或多个比较之间是否独立，可以利用系数公式来方便地确定。如对前一组对比的判断如下：

C_1		C_2		C_3
1		0		−1
0		1		−1
$\frac{1}{2}$		$\frac{1}{2}$		−1
0	+	0	+	（−1）＝−1

心理学研究方法

VISIT…

LANZAROTE Caliente.com

如果我们将比较中每一列系数相乘，然后再将乘积相加，结果可以帮助我们判断对比的正交性问题。当得到的相加结果不是 0 的时候，我们可以知道这组比较之间是非正交的，这意味着几个比较提出的问题是相互不独立的。在第一例中，我们可以看出这是一组非正交的对比。

我们对后一组对比的判断如下：

C_1		C_2		C_3
$\frac{1}{2}$		$\frac{1}{2}$		-1
1		-1		0
$\frac{1}{2}$	$+$	$\left(-\frac{1}{2}\right)$	$+$	$0=0$

这时，当我们将比较中每一列系数相乘，然后将乘积再相加，得到的结果是 0 的时候，我们可以知道两个比较是正交的。这意味着两个比较提问的是相互独立的问题。相互独立的问题指，当我们知道一个问题的结论时，并不知道对另一个问题的结论。或者说，两个比较的正交性或独立性表示知道第一个对比是显著的，不能预期第二个对比是否显著。在第二例中，我们可以看出这是一组正交对比。

我们还可以用公式表示两个比较是正交的：

$$\sum C_{ai}C_{bi}=0$$

即当两个或多个比较是正交的，那么它们的系数乘积的总和应当等于 0。

（二）正交对比的数量

对于一组有 p 个水平的处理平均数，应当有几个正交对比呢？实际上，在 p 个平均数之间，不能超过 $p-1$ 个正交的对比，或者说正交对比的数量与处理的自由度相同，即 $p-1$。假如对于一个三个平均数的实验，可能的比较数量为六。而在六个可能的比较中，一组互相正交的对比只有两个，共有三组正交的比较，它们可以用系数的方式表示如下：

第一组：	1	$-\frac{1}{2}$	$-\frac{1}{2}$
	0	1	-1
第二组：	$-\frac{1}{2}$	1	$-\frac{1}{2}$

$$\begin{array}{ccc} 1 & 0 & -1 \end{array}$$

第三组：

$$\begin{array}{ccc} -\frac{1}{2} & -\frac{1}{2} & 1 \\ 1 & -1 & 0 \end{array}$$

研究者可能不一定对所有的 $p-1$ 个正交对比感兴趣。如果我们考虑前面提到的生字密度对阅读理解影响的研究，其中三种生字密度水平分别是：5∶1（A1），10∶1（A2），20∶1（A3），那么，以上的第二组正交对比就是没有意义的。因为它提出的问题是：儿童对生字密度为 5∶1 和 20∶1 的文章的阅读理解是否与生字密度为 10∶1 的文章的阅读理解有差别，这个问题在实践中可能是没有意义的。

当实验中有四个处理平均数时，一组相互正交的对比有三个，共有四组正交的比较。它们可以用系数的方式表示如下：

第一组：

$$\begin{array}{cccc} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{array}$$

第二组：

$$\begin{array}{cccc} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{array}$$

第三组：

$$\begin{array}{cccc} 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{array}$$

第四组：

$$\begin{array}{cccc} 1 & -1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & -1 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -1 \end{array}$$

（三）正交对比的独立性

我们在前面曾经提到，非正交对比是指两个比较的问题是不相互独立的，而正交对比是指两个比较的问题是相互独立的。下面我们将从处理平

方和的角度再看一看正交对比的独立性。假如在一个实验中，A 因素的主效应是显著的。A 因素有三个处理水平，得到三个处理平均数分别是 16、7 和 11。每组 5 个被试。进行多重比较时，各个比较的平方和与 A 因素处理平方和的关系是：

$$SS_{\mathrm{A}}=\sum SS_{\mathrm{Acomp}}$$

其中 SS_{Acomp} 为每个比较的平方和，在我们的例子里，可以写做：

$$SS_{\mathrm{A}}=SS_{\mathrm{Acomp1}}+SS_{\mathrm{Acomp2}}$$

如果我们对第一组与第二组、第三组的差异是否显著，以及第二组与第三组之间差异是否显著感兴趣，可以进行一组正交的比较（见表 10-2）。

表 10-2　三个处理水平的事先正交对比

	A1	A2	A3
平均数	16	7	11
被试数	5	5	5
总和	80	35	55
比较 1	1	$-\frac{1}{2}$	$-\frac{1}{2}$
比较 2	0	1	-1

利用各处理水平的总和计算平方和如下。

根据总和计算的公式：

$$SS_{\mathrm{Acomp}}=\frac{\left(\sum_{j=1}^{p}C_jA_j\right)^2}{n_j\sum_{j=1}^{p}C_j^2}$$

其中 C_j 是比较系数，A_j 是参加比较的各处理水平的总和，p 为自变量的水平数，n_j 是各组被试的数量。

$$SS_{\mathrm{Acomp1}}=\frac{\left[1\times 80+\left(-\frac{1}{2}\right)\times 35+\left(-\frac{1}{2}\right)\times 55\right]^2}{5\times\left[1^2+\left(-\frac{1}{2}\right)^2+\left(-\frac{1}{2}\right)^2\right]}=163.33$$

第十章 多重比较：对比

$$SS_{Acomp2}=\frac{[0\times80+1\times35+(-1)\times55]^2}{5\times[0^2+1^2+(-1)^2]}=40$$

表 10-3 事先正交对比的方差分析表

来源	平方和	自由度	均方	F
处理 A	203.33	2		
比较 1	163.33	1	163.33	29.70***
比较 2	40	1	40	7.27*
误差	66.00	12	5.50	
合计	269.33	14		

这时，可以看到，当一组比较是正交时，SS_A 正好分解为比较 1 的平方和 SS_{Acomp1} 和比较 2 的平方和 SS_{Acomp2} 相加的和（见表 10-3）。然而，如果我们感兴趣于第一组与第二组的差异是否显著，以及第一组与第三组之间差异是否显著，这时进行的是一组非正交的比较（见表 10-4）。

表 10-4 三个处理水平的事先非正交对比

	A1	A2	A3
平均数	16	7	11
被试数	5	5	5
总和	80	35	55
比较 1	1	−1	0
比较 2	1	0	−1

利用各处理水平的总和计算平方和如下。

$$SS_{Acomp1}=\frac{[1\times80+(-1)\times35+0\times55]^2}{5\times[1^2+(-1)^2+0^2]}=202.5$$

$$SS_{Acomp2}=\frac{[1\times80+0\times35+(-1)\times55]^2}{5\times[1^2+0^2+(-1)^2]}=62.5$$

这时，会看到，SS_{Acomp1} 和 SS_{Acomp2} 相加的和会大于 SS_A，即：

$$202.5+62.5=265.0$$

表 10-5　事先非正交对比的方差分析表

来源	平方和	自由度	均方	F
处理 A	203.33	2		
比较 1	202.5	1	202.5	36.82***
比较 2	62.5	1	62.5	11.36**
误差	66.00	12	5.50	
合计	269.33	14		

从方差分析表中可以看出两个非正交比较之间有变异的重叠。在一个研究中，选择哪种比较主要取决于我们要探讨的问题。当我们选择进行一组正交比较，这组正交比较中应当含有我们感兴趣的某个或某些比较，但不一定是所有的比较。另外，在进行实验设计时，并不是必须进行正交对比分析的设计。例如，当我们感兴趣于几个实验组各自与控制组相比较的差异时，或在启动实验中我们感兴趣于各个启动条件与控制条件的差异所得到的启动效应时，所进行的对比往往是非正交的。然而，这是研究者的理论兴趣所在，即使是非正交的对比，研究者也只能选择这样的比较。

与非正交对比相比，正交对比的使用有以下几个优点。第一，每个对比的解释是完全独立于与其正交的其他对比的，因此在对一组问题的回答中，对一个问题的回答不受其他问题的影响。第二，由于正交对比中对比之间互相独立，因此它是用最小数量的对比抽取最大量的有用信息，信息之间没有重叠或冗余。如果有 $p-1$ 个独立信息，$p-1$ 个正交对比使用了所有的信息，而非正交对比需要多于 $p-1$ 个对比所使用的信息。例如，在语音、词素和传统教学法的研究中，当研究者关心“实验教学法与传统教学法是否有差别，语音教学法与词素教学法之间是否有差别”时，或关心“语音教学法与传统教学法是否有差别，语素教学法与传统教学法是否有差别，以及实验教学法与传统教学法是否有差别”这些问题时，两组比较所回答的问题是类似的，然而前者是正交对比，可以用更少的比较回答类似的问题。

全方差分析的基本思想是，如果全方差分析显著，至少其中一个比较检验是显著的。全方差分析 F 检验的公式：

$$F = MS_A / MS_E$$

我们进行一组正交对比，就像进行一个全方差分析。一组正交对比的平方和 $\sum SS_{Acomp}$ 与全方差分析的处理平方和 SS_A 是相同的。一组正交对比的自由度（$p-1$），与一个全方差分析的自由度（$p-1$）是相同的。因此，如果一组正交对比检验显著，则方差分析肯定是显著的。从这个角度说，全方差分析相当于一组正交比较。

$$F = \sum SSC_j / MS_E$$

但在一组非正交对比中，当 $\sum SS_{Acomp} > SS_A$ 时，可能出现至少一个对比显著，而全方差分析不显著的情况。这种情况下，这个对比的显著性可能是不真实的，或者说可能与累积错误有关。因此，当我们进行非正交对比时，更安全的方法是首先进行全方差分析显著性检验，然后进行平均数的多重比较检验，找出其中一个或多个显著的平均数之间的差异。

（四）正交对比：趋势分析

当实验中自变量和因变量均为定量的变量时，有时自变量的变化和因变量的变化呈现出一定的规律性。随着自变量水平的量的增加，处理效应平均数稳定地上升或下降的现象被称为实验结果的线性变化。而如果随着自变量水平的量的增加，处理效应平均数的变化是多方向的，则可能反映了实验处理效应的非线性变化趋势。趋势分析（trend analysis）主要指对研究者关于实验处理效应为线性或非线性趋势理论假设的正确性的检验。

下面我们举例说明对比在实验研究结果的趋势分析中的应用。为了计算方便，我们可以将对比的系数整数化。例如，一个对比$\left(1,\ -\frac{1}{2},\ -\frac{1}{2}\right)$可以表示为（2，−1，−1），这两个对比提出的假说是同样的。假设我们想要探讨工作压力与工作成绩的关系。我们选取四组精细零件生产工人，告诉所有的工人每生产 100 件零件，可接受的报废数量为 10 件。但给四组工人的另一部分指导语是不同的。告诉第一组工人，他们不会因为超过报废数量的标准而被扣工资；告诉第二组工人，如果超过报废数量标准就要扣掉四分之一的工资；告诉第三组工人，如果超过报废数量标准就要扣掉

一半工资；告诉第四组工人，如果超过标准就扣掉全部的工资。可见，随着对工人由于超过报废率所扣工资比率的提高，他们工作的压力水平也就越大。我们用最后的各组生产合格率作为衡量的指标。在这个领域的前人研究中已提出三种理论。一种理论认为随着压力水平的不断增加，工作的成绩会不断地下降，因此认为第一组工人的合格率最高，其余组随着扣工资比率的提高，合格率会随之降低。这种理论预期实验处理效应为线性变化。另一种理论认为，随着压力的不断增加，工作的成绩会不断地提高，那么第一组的合格率最低，其余组的合格率会随着压力的增加不断提高。这种理论也预期实验处理效应为线性变化。还有一种理论认为，压力的增加在开始阶段可以促进工作成绩，但是当压力提高到一定程度后，就会降低工作的成绩，也就是说，第二组的合格率应该高于第一组，第三组可能和第二组差不多，而第四组可能因为其压力过大，合格率会低于第三组。可以看到，第三种理论预期实验处理效应为非线性变化。

如何使用趋势分析的方法检验理论假设的正确性？首先，我们需要将理论假设的语言表达形式转换成数量化形式的对比系数。有许多方法可以将这些语言的描述翻译成更加数量化形式的对比系数，一个很有用的途径是按以下四个步骤选择系数：

（1）对每种处理条件下产生理想的因变量观测平均数值，形成理论预期的结果模式；

（2）预计这些理想值的平均数值的总平均数，即预期的总平均；

（3）将每种预期条件下的平均数减去预期的总平均数；

（4）简化数字，以便进行最简单的计算。

以上述实验为例，三种理论对实验处理效应的预期可以表示成以下的系数矩阵（见表 10-6），其中第一个对比（A）表明了生产合格率随着工作压力增加而线性下降的趋势，第二个对比（B）表明了生产合格率随压力增加线性上升的趋势，而第三个对比（C）表明了合格率随压力增加而先上升、后下降的非线性趋势。这三个比较将能够检验工人实际生产零件合格率符合哪种理论预期。需要注意的是每个比较中的系数之和为 0。

表 10-6　三种理论预期的比较系数

	无惩罚	扣四分之一工资	扣一半工资	扣全部工资
线性下降（A）	+3	+1	−1	−3
线性上升（B）	−3	−1	+1	+3
非线性（C）	−1	+1	+1	−1

在实验中，第一步，我们可以收集在不同压力条件下工人的产品合格数量。表 10-7 中是每种条件下五位工人生产合格产品的数量。

表 10-7　四组工人产品合格数量的原始数据（单位：件）

实验组			
无惩罚	扣四分之一工资	扣一半工资	扣全部工资
79	91	97	90
85	95	91	80
70	96	90	85
91	92	90	83
86	89	92	82

第二步，我们可以得到四种条件下工人的平均合格率、总数量，并与三组比较的系数相结合（见表 10-8）。

表 10-8　四组工人的平均合格率及比较系数

	实验组			
	无惩罚	扣四分之一工资	扣一半工资	扣全部工资
平均合格率	82.2	92.6	92	84
人数	5	5	5	5
总数量	411	463	460	420
线性下降（A）	+3	+1	−1	−3
线性上升（B）	−3	−1	+1	+3
非线性（C）	−1	+1	+1	−1

第三步，对实验结果进行对比的检验。如果已知全方差分析的误差均方值，可以用 t 检验的方法，直接采用以下公式：

$$t=\frac{\sum_{j=1}^{p}C_j\overline{X_j}}{\sqrt{MS_{err}\sum_{j=1}^{p}\frac{C_j^2}{n_j}}}$$

具体计算如下：

$$t_A=\frac{3\times82.2+1\times92.6+(-1)\times92+(-3)\times84}{\sqrt{24\times\left[\frac{3^2}{5}+\frac{1^2}{5}+\frac{(-1)^2}{5}+\frac{(-3)^2}{5}\right]}}=-0.49$$

$$t_B=\frac{(-3)\times82.2+(-1)\times92.6+1\times92+3\times84}{\sqrt{24\times\left[\frac{(-3)^2}{5}+\frac{(-1)^2}{5}+\frac{1^2}{5}+\frac{3^2}{5}\right]}}=0.49$$

$$t_C=\frac{(-1)\times82.2+1\times92.6+1\times92+(-1)\times84}{\sqrt{24\times\left[\frac{(-1)^2}{5}+\frac{1^2}{5}+\frac{1^2}{5}+\frac{(-1)^2}{5}\right]}}=4.199$$

查 t 值表表明，只有第三个对比 t_C 达到了 $p<0.001$ 的差异统计显著性，说明实验结果支持第三种理论，或者说趋势分析表明实验处理效应是非线性变化的。

我们也可以用 F 检验的方法进行趋势分析。利用表 10-8 中各处理水平的总和计算平方和，使用以下公式：

$$SS_{AcompA}=\frac{[3\times411+1\times463+(-1)\times460+(-3)\times420]^2}{5\times[3^2+1^2+(-1)^2+(-3)^2]}=5.76$$

$$SS_{AcompB}=\frac{[(-3)\times411+(-1)\times463+1\times460+3\times420]^2}{5\times[(-3)^2+(-1)^2+1^2+3^2]}=5.76$$

$$SS_{AcompC}=\frac{[(-1)\times411+1\times463+1\times460+(-1)\times420]^2}{5\times[(-1)^2+1^2+1^2+(-1)^2]}=423.2$$

表 10-9　趋势分析的方差分析表

来源	平方和	自由度	均方	F
处理（工作压力）	432.2	3	144.07	6.00
线性下降（A）	5.76	1	5.76	0.24
线性上升（B）	5.76	1	5.76	0.24
非线性（C）	423.2	1	423.2	17.63**
误差	384.0	16	24	

查 F 值表表明，只有第三个对比 SS_{AcompC} 达到了 $p=0.001$ 的差异统计显著性（见表 10-9）。F 检验的结论与 t 检验的结论是一致的。

第三节　计划的或事先比较的 SPSS 常用计算方法

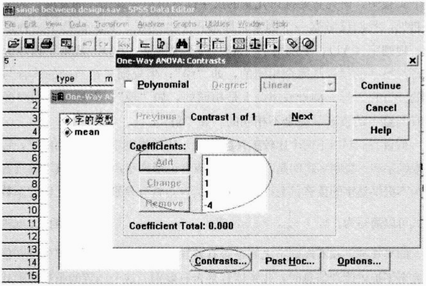
一、事先非正交对比的 SPSS 操作

我们在前两节中已经介绍了计划的事先比较的计算公式，包括进行正交和非正交对比的检验方法。在本节中，我们将介绍如何使用 SPSS 软件进行计划的事先比较的显著性检验。事先的对比主要使用的是 t 检验公式：

$$t=\frac{\sum_{j=1}^{p}C_j\overline{X}_j}{\sqrt{MS_{\text{err}}\sum_{j=1}^{p}\frac{C_j^2}{n_j}}}$$

Contrasts 在 SPSS 中也有相应的窗体操作，但是只有在 One-Way 里面其操作更为灵活方便。下面举例说明非正交对比的计算。事先非正交对比指对比是研究者在实验开始之前根据理论或前人文献确定的一组比较，所设定的几个对比之间是相互不独立的。

例如，在一个儿童延迟抄写的研究中，研究者要探讨小学儿童的汉字正字法意识的发展。实验中的汉字自变量有五种水平：（1）熟悉的独体字；（2）熟悉的合体字；（3）不熟悉的合体字，但字的部件是熟悉的；（4）不熟悉的合体字，字的部件也是不熟悉的；（5）随机笔画组成的图形。如果我们感兴趣于考察儿童在延迟抄写真字与随机笔画组成的图形之间的差别，可以比较延迟抄写前四种字和延迟抄写第五种图形的成绩之间的差异。在 SPSS 中的操作见下页图，在 Coefficients 后面的框里依次按序填上每种条件对应的权重系数。在这里我们要比较前面四种条件与第五种的差异，所以前面四个的系数都设为 1，第五个设为－4。填好后，按“Continue”，然后就可以运行了。



我们来看看结果输出，表 10-10 中 Contrast Coefficients 部分显示的是上图输入的对比系数，可以检查一下是否正确。表 10-10 中 Contrast Tests 部分是对这一个对比的检验，前面说了它用到的是 t 检验。可以看到，延迟抄写前四种字的成绩与延迟抄写第五种图形的成绩确实是有明显差异的。

表 10-10　事先非正交对比检验的结果输出

Contrast Coefficients

Contrast	字的类型				
	1.00	2.00	3.00	4.00	5.00
1	1	1	1	1	-4

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
MEAN	Assume equal variances	1	1.0200	0.2024	5.040	30	0.000
	Does not assume equal	1	1.0200	0.2873	3.550	6.861	0.010

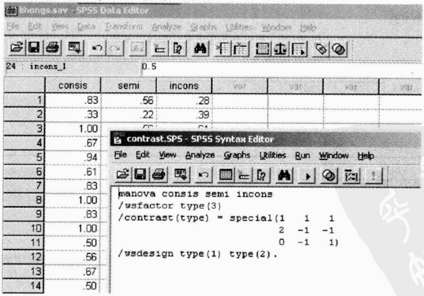
二、事先正交对比的 SPSS 操作

我们再以另一组实验数据为例，看看事先的正交对比在 SPSS 中如何实现，同时我们也比较一下事先计划的比较和事后的多重比较的差异。在

这个实验中，研究者要探讨儿童的汉字读音中声旁规则性的发展。为了简单起见，我们仅考察一个因素，即形声字的类型。实验设计了三种形声字：规则字（A1）、半规则字（A2）和不规则字（A3）。规则字指声旁的读音与整字的读音是完全相同的，如“纷”；半规则字指声旁的读音与整字的读音在声母、韵母或声调上是相同的，如“现”；不规则字指声旁的读音与整字的读音是完全不相同的，如“猜”。

假设我们对以下两个比较感兴趣。(1) 比较规则字和无规则的字，包括半规则字和不规则字之间是否有差异，或者说形声字声旁提供整字的读音信息与不提供整字的读音信息时，儿童的读音是否有差别。这个比较的系数的形式可以简写为：1，$-\frac{1}{2}$，$-\frac{1}{2}$，或 2，−1，−1。(2) 比较两种无规则的字，即半规则字和不规则字之间是否有差异，或者说声旁提供部分读音信息与完全不提供读音信息时，儿童读音是否有差别。这个比较的系数的形式可以简写为：1，−1。当同时关心这两个对比时，这是一组正交的比较。

首先，在 SPSS 中输入数据，并按实验设计排列好。字的类型在这里是被试内因素。这里的被试内因素的 Contrast 需要在 Syntax 窗体中编写。数据格式及 Syntax 句法如下图。



结果输出如下（见表 10-11）。

表 10-11 事先正交对比检验的结果输出

```
Tests of Between-Subjects Effects.

Tests of Significance for T1 using UNIQUE sums of squares
Source of Variation          SS      DF      MS        F  Sig of F

WITHIN+RESIDUAL            7.78      60     .13
CONSTANT                  46.59       1   46.59   359.13      .000

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Estimates for T1
--- Individual univariate .9500 confidence intervals

CONSTANT

 Parameter     Coeff.  Std. Err.    t-Value     Sig.  t Lower -95%  CL- Upper

      1  .873912703     .04611   18.95083    .00000     .78167      .96616

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

* * * * * * A n a l y s i s   o f   V a r i a n c e -- design   1 * * * * * *

Tests involving 'TYPE(1)' Within-Subject Effect.

Tests of Significance for T2 using UNIQUE sums of squares
Source of Variation          SS      DF      MS        F  Sig of F

WITHIN+RESIDUAL            2.36      60     .04
TYPE(1)                    3.35       1    3.35    84.99      .000

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Estimates for T2
--- Individual univariate .9500 confidence intervals

TYPE(1)
```

Parameter	Coeff.	Std. Err.	t-Value	Sig. t	Lower -95%	CL- Upper
1	.234240822	.02541	9.21896	.00000	.18342	.28507

\- -

* * * * * * A n a l y s i s   o f   V a r i a n c e -- design 1 * * * * * *

Tests involving 'TYPE(2)' Within-Subject Effect.

Tests of Significance for T3 using UNIQUE sums of squares

Source of Variation	SS	DF	MS	F	Sig of F
WITHIN+RESIDUAL	1.11	60	.02		
TYPE(2)	.20	1	.20	10.62	.002

\- -

Estimates for T3

--- Individual univariate .9500 confidence intervals

TYPE(2)

Parameter	Coeff.	Std. Err.	t-Value	Sig. t	Lower -95%	CL- Upper
1	-.05667158	.01739	-3.25926	.00184	-.09145	-.02189

\- -

从输出中我们可以看到两个比较的结果。其中，比较（1），即 type(1)，是指规则字与无规则的字（即半规则字和不规则字）的差异比较，对应于句法中的 2，－1，－1。其检验结果是：$F(1, 60)=84.99$，$p=0.000$，说明规则字与无规则的字的成绩差异在统计上是显著的。比较(2)，即 type(2)，是指两种无规则的字之间的差异比较，对应于句法中的 0，－1，1。其检验结果是：$F(1, 60)=10.62$，$p=0.002$，说明同样半规则字和不规则字的差异也是统计上显著的。总之，这一组正交对比的结果表明，儿童读规则字的正确率高于读没有规则的字的正确率，读半规则字的正确率高于读不规则字的正确率。结果细致地揭示了儿童的汉字读

音中声旁规则性的作用。

本章主要观点

• 当全方差分析中的主效应差异显著时，需要进行额外的检验，即多重比较。然而额外的多重比较检验带来的一个重要问题是Ⅰ型错误的增加。

• 在多重比较检验中，我们需要区分每个比较的Ⅰ型错误率和每个实验的错误率或实验错误率。如果研究者在一个实验中需要进行多个平均数的比较，整个实验犯Ⅰ型错误的可能性就是实验错误率。

• 完成多重比较检验的技术包括计划的或事先的正交对比、事先的非正交对比和事后的非正交比较等。计划的或事先的比较是指比较的确定是基于理论或前人经验的，是在实验数据收集之前确定的。无计划或事后的比较是指实验数据收集后，并看到结果之后确定要进行的比较，实验者在实验实施前没有特殊的假设。

• 当研究者对同一组数据提出两个或两个以上关于比较的问题时，还需要了解这两个问题之间是否相互独立。评价比较之间是否独立的问题就是对比的正交性问题。非正交对比是指两个比较的问题是不相互独立的。正交对比是指两个比较的问题是相互独立的。

• 对比是一个在实验研究和数据分析中非常有用的工具。对比分析中，比较是基于理论假说或前人的研究结果的。它使研究者可以去检验一个或几个高度特定的假说。对比是特别有效的发现和解释数据意义的工具。

• 随着自变量水平的量的增加，处理效应平均数稳定地上升或下降的现象被称为实验结果的线性变化。而如果随着自变量水平的量的增加，处理效应平均数的变化是多方向的，则可能反映了实验处理效应的非线性变化趋势。趋势分析主要指对研究者关于实验处理效应为线性或非线性趋势理论假设的正确性的检验。

思考题

1. 什么是多重比较？为什么要进行多重比较检验？多重比较给推论

统计带来的问题是什么？

2. 什么是多重比较中的累积错误？什么是每个比较的错误率，什么是实验错误率，这两者的关系是什么？

3. 多重比较有哪些种类？不同种类的比较通常使用哪些检验方法？

4. 什么是事先比较？在什么情况下使用事先比较？

5. 什么是正交比较？正交比较有什么优点？如何判断一组比较是不是正交的？

第十一章
多重比较：事后比较

无计划或事后比较是在方差分析 F 检验之后，根据数据结果模式，对感兴趣的平均数之间差异进行评估，而研究者在实验实施前没有特殊的假设。例如，从实验结果中看到 A1、A2 和 A3 水平的平均数有差异，我们想知道它们之间哪些差异是显著的，这就是事后比较。事后比较的主要形式是进行所有可能的处理平均数的成对比较，目的是希望最大限度地获得研究数据中的有用信息。事后比较通常需要进行大量、穷尽的成对比较，实验处理中的每个平均数都会被包含进多个与其他平均数的比较。因而，事后比较同样面临的一个重要问题是：实验错误率会随着比较的数量增加而增加。

我们想象设计一个实验，事先没有有关检验平均数之间差异的假说，而希望有一种方法能找出可能导致处理主效应显著的来源，最好的方法是穷尽所有可能的成对的平均数差异检验。由于比较数量的增加会带来更大的犯Ⅰ型错误的可能性，我们需要使用一些方法技术来解决Ⅰ型错误累积的问题。所有事后检验技术都是通过减小拒绝区域来解决这个问题。其基本想法是，如果减小拒绝区域，我们更难以拒绝虚无假说，就会犯较少的Ⅰ型错误，实验错误率就会下降。

第一节　几种事后成对比较方法

当我们对一个实验处理的主效应没有事先的理论假设时，最好的办法是检验所有可能的平均数之间的差异。这增加了发现有意义的差异的可能性，但也增加了犯Ⅰ型错误的可能性。统计学家已经发展了各种检验和统计方法，对多重比较中的累积错误进行校正。有三条解决问题的途径：一

个极端的途径是不考虑实验错误率，或者说对统计结果不进行任何校正，最小显著差异检验属于这种性质；另一个极端的途径是进行非常严格的校正，Scheffe 检验是这种性质的，这种方法通过有效减小每个比较的错误率，以控制实验错误率。处于两者中间的途径则是调整的大小依赖于被比较的平均数的数量，Tukey-HSD 检验和 Newman-Keuls 检验属于这一类。

一、几种常用的事后比较检验方法

（一）最小显著差异检验

最小显著差异检验（least significant difference test，简称 LSD 检验），其特点是对全方差分析显著后所作的比较不作任何校正。进行 LSD 检验的前提是要求全方差分析显著，或虚无假说被拒绝，然后进行包含若干组无特殊假设的事后比较。如果全方差分析不显著，不能进行进一步的多重比较检验。进行成对比较的检验方法可以使用多重 t 检验，任何一个比较的显著性可以由以下公式计算。但需要注意的是，当方差齐性检验表明各组误差变异不同质的时候，不能进行此检验。

当参加比较的两组被试相等时，可以使用的公式如下：

$$t=\frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{2MS_{err}}{n}}}$$

其中$\overline{X_1}$、$\overline{X_2}$是参加比较的平均数，MS_{err}是全方差分析中的误差项，n 是每组的被试数。

当两组被试不相等时，可以使用的公式如下：

$$t=\frac{\overline{X_1}-\overline{X_2}}{\sqrt{MS_{err}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

其中 n_1、n_2 是两个不等组的被试数。

还可以用一种更简单、直接的方法，直接计算出两个平均数的差，然后查临界值表：

$$\text{LSD}=t\sqrt{\frac{2MS_{err}}{n}}$$

LSD 检验对实验错误率不作任何控制，或者说将每个比较的错误率

和实验错误率等同起来，忽略累积错误。它使用与计划的事先比较同样的检验方法进行事后比较的检验。唯一不同的是：LSD 检验需要在全方差分析的 F 检验显著后进行，而事先计划的比较则不一定必须检验全方差分析是否显著。

还有一些对实验错误率调整的比较检验方法，尝试对累积错误进行校正。Tukey-HSD 检验、Newman-Keuls 检验和 Scheffe 检验都属于这一类。

（二）Tukey-HSD 检验

Tukey-HSD 检验是一种 q 检验。它的基本假定是，如果有 j 个独立的样本，每个样本有一个平均数 $\overline{X_j}$，将这些平均数按大小排列。进行成对比较的检验方法可以使用多重 q 检验。

当参加比较的两组被试相等时，q 检验使用的公式如下：

$$q=\frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{MS_{err}}{n}}}$$

其中 $\overline{X_1}$ 通常指参加比较中数量较大的平均数，$\overline{X_2}$ 通常指参加比较中数量较小的平均数，MS_{err} 是全方差分析中的误差项，n 是每组的被试数。

当参加比较的两组被试不相等时，可以使用的公式如下：

$$q=\frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{MS_{err}}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

其中，MS_{err} 是全方差分析中得到的误差项。当各组被试数相等时，n 是样本数。当各组被试数不相等时，可以用 $\frac{2(n_1 \times n_2)}{n_1+n_2}$ 代替 n，其中 n_1 是平均数较大组的样本数，n_2 是平均数较小组的样本数。

Tukey-HSD 检验也可以用一种更简单、直接的方法，即直接计算出两个平均数的差，然后查临界值表。

当参加比较的两组被试相等时，使用的公式如下：

$$HSD_{\overline{X_1}-\overline{X_2}}=q\sqrt{\frac{MS_{err}}{n}}$$

当两组被试不相等时，使用的公式如下：

$$HSD_{\overline{X_1}-\overline{X_2}}=q\sqrt{\frac{MS_{err}}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

当我们检验所有成对平均数之间的差异时，可以直接查 q 分布的临界值表。如果一对平均数的绝对差异（$\overline{X_1}-\overline{X_2}$）等于或大于 HSD，我们就可以拒绝平均数相等的假说。用这种方法，进行多个比较和一个比较是相同的，也就是说当我们查表中的 q 值时，错误率是等于我们所用的 α 的。在 Tukey-HSD 检验中，有两个重要的参数：参加比较的组数和 MS_{err} 的自由度。

下面我们举一个例子来说明 Tukey-HSD 检验的过程。研究者要探讨文章生字密度对学生阅读理解的影响。这是一个单因素完全随机实验。全方差分析表明，阅读理解随着文章中生字密度的增加而下降，即生字密度的主效应是显著的。研究者进一步感兴趣于哪些生字密度不同［如生字密度 5∶1(A1)、10∶1(A2)、15∶1(A3) 和 20∶1(A4)］的文章的阅读理解分数之间存在差异。研究者在实验前对四种生字密度可能对学生阅读理解产生的影响没有特别的假说，而是在方差分析中发现生字密度的效应非常显著，因而需要进一步检验全部可能的成对平均数之间的差异。要检验的假说是：

(1) H_0: $\mu_1-\mu_2=0$

(2) H_0: $\mu_1-\mu_3=0$

(3) H_0: $\mu_1-\mu_4=0$

(4) H_0: $\mu_2-\mu_3=0$

(5) H_0: $\mu_2-\mu_4=0$

(6) H_0: $\mu_3-\mu_4=0$

从全方差分析中，我们已知组内均方和（$MS_{err}=2.813$）、自由度（$df=28$）、组的数目（$p=4$）和每组的被试数（$n=8$）等信息。当方差分析表明生字密度的主效应显著后，可首先将各处理水平的平均观测值按从小到大排列等级（见表 11-1）。

表 11-1　平均数等级排列表

等级	1	2	3	4
平均数	$\overline{X_2}=3.88$	$\overline{X_1}=4.38$	$\overline{X_3}=7.00$	$\overline{X_4}=10.00$

然后利用两两平均数之间的差数作表。

表 11-2 平均数之间差数的绝对值及 q 检验的显著性

	$\overline{X_2}$ (3.88)	$\overline{X_1}$ (4.38)	$\overline{X_3}$ (7.00)	$\overline{X_4}$ (10.00)
$\overline{X_2}$		0.50	3.12**	6.12**
$\overline{X_1}$			2.62	5.62**
$\overline{X_3}$				3.00**

下一步，我们可以直接用各对平均数的差异与查表所得的临界值进行比较。查 q 值表得到，当 $df=28$，$p=4$ 时，$q_{0.01}$ 是 4.80，通过公式 $q_{0.01}\sqrt{\frac{MS_{err}}{n}}=2.842$ 的计算，可得到 HSD 的临界值是 2.842。从表 11-2 中可以看到，差值大于临界值 2.842 的有四组，即$\overline{X_2}$与$\overline{X_3}$，$\overline{X_2}$与$\overline{X_4}$，$\overline{X_1}$与$\overline{X_4}$，$\overline{X_3}$与$\overline{X_4}$，即这四对平均数之间的比较是在 0.01 水平显著的。

需要注意的是，Tukey-HSD 检验对所有可能的比较使用了单一的临界值，检验的目的是对所有可能的比较控制实验错误率。它将所有的比较看成是相同的，即对所有的比较使用相同的调整方式。

（三）Newman-Keuls 检验

Newman-Keuls（SNK）检验使用与 Tukey-HSD 检验同样的技术进行事后多重比较，即先将平均数按大小排列，计算各对平均数之间的差异。但与 Tukey-HSD 检验不同的是，它不是用同样的标准检验所有的差异，而是采用阶梯方法（stairstep）。Newman-Keuls 检验假定各组平均数是按大小排列的，为了进行所有可能的成对比较，先将相邻的平均数进行比较，然后将相隔一个的平均数进行比较，再将相隔两个的平均数进行比较，最后比较最大与最小的平均数。不同阶梯上各个比较之间的差异检验会在不同的 α 水平上进行。因此，与 Tukey-HSD 检验相比，Newman-Keuls 检验更容易拒绝虚无假说。

Newman-Keuls 检验的重要特点是通过调整将每一个比较的Ⅰ型错误置于特定的 α 水平。当 $r=2$ 时，没有校正；当 $r>2$ 时，使用校正。通过校正，使得每个比较，甚至极端组的比较，都被调整在 α 水平。这在很大

程度上减小了错误率。

我们继续使用上面的例子来说明。首先将各处理水平的平均观测值按从小到大排列等级。

表 11-3 平均数等级排列表

比较等级（r）	1	2	3	4
平均数	$\overline{X_2}=3.88$	$\overline{X_1}=4.38$	$\overline{X_3}=7.00$	$\overline{X_4}=10.00$

然后利用两两平均数之间的差数作表。

表 11-4 平均数之间差数的绝对值及 q 检验的显著性

	$\overline{X_2}$ (3.88)	$\overline{X_1}$ (4.38)	$\overline{X_3}$ (7.00)	$\overline{X_4}$ (10.00)
$\overline{X_2}$		0.50	3.12**	6.12**
$\overline{X_1}$			2.62**	5.62**
$\overline{X_3}$				3.00**

根据上面介绍的 q 检验公式：

$$\overline{X_1}-\overline{X_2}=q\sqrt{\frac{MS_{err}}{n}}$$

可以求出各等级的两两平均数相比较时的 0.01 水平显著的临界值。在 Newman-Keuls 检验中，也有两个重要的参数：比较等级 r 和 MS_{err} 的自由度。比较等级 r 指在平均数等级排列表（见表 11-3）中两个平均数相邻的情况。当两个平均数相邻时，如将$\overline{X_2}$与$\overline{X_1}$，$\overline{X_1}$与$\overline{X_3}$，$\overline{X_3}$与$\overline{X_4}$相比较时，$r=2$。当两个平均数中间相隔一个平均数时，如将$\overline{X_2}$与$\overline{X_3}$，$\overline{X_1}$与$\overline{X_4}$相比较时，$r=3$。当两个平均数中间相隔两个平均数时，如将$\overline{X_2}$与$\overline{X_4}$相比较时，$r=4$。根据比较等级 r 和误差平方和的自由度（$df=28$），查表 q 可知：

$r=2$ 时，$q_{0.01}=3.89$；

$r=3$ 时，$q_{0.01}=4.45$；

$r=4$ 时，$q_{0.01}=4.80$。

进一步可以算出各等级的两两平均数相比较时的 0.01 水平显著的临

界值：

$r=2$ 时，$q_{0.01}\sqrt{\dfrac{MS_{err}}{n}}=2.307$；

$r=3$ 时，$q_{0.01}\sqrt{\dfrac{MS_{err}}{n}}=2.639$；

$r=4$ 时，$q_{0.01}\sqrt{\dfrac{MS_{err}}{n}}=2.842$。

将各对平均数差异与相应比较等级的临界值相比，如果两个平均数之间差异大于该临界值，则该平均数比较是统计显著的。在本例子中可以看到：

当 $r=2$ 时，其临界值为 2.307，则

$\overline{X_1}-\overline{X_2}=0.05$　平均数之间差异小于临界值（$0.05<2.307$）

$\overline{X_3}-\overline{X_1}=2.62$　平均数之间差异大于临界值（$2.62>2.307$）

$\overline{X_4}-\overline{X_3}=3.00$　平均数之间差异大于临界值（$3.00>2.307$）

当 $r=3$ 时，其临界值为 2.639，则

$\overline{X_3}-\overline{X_2}=3.12$　平均数之间差异大于临界值（$3.12>2.639$）

$\overline{X_4}-\overline{X_1}=5.62$　平均数之间差异大于临界值（$5.62>2.639$）

当 $r=4$ 时，其临界值为 2.842，则

$\overline{X_4}-\overline{X_2}=6.12$　平均数之间差异大于临界值（$6.12>2.842$）

从 Newman-Keuls 检验结果中可以看出，除了$\overline{X_1}$与$\overline{X_2}$外，其余五对平均数的比较均是统计显著的（$p<0.01$）。如果我们将 Newman-Keuls 检验与 Tukey-HSD 检验的结果相比较会发现，对同样的一组数据，Tukey-HSD 检验中有四对平均数之间的比较差异显著，而 Newman-Keuls 检验中有五对平均数之间的比较差异显著。因此可以说，与 Tukey-HSD 检验相比，Newman-Keuls 检验更容易拒绝虚无假说。Tukey-HSD 检验使用了更严格的检验标准。

（四）Scheffe 检验

Scheffe 检验的前提是全方差分析显著。当全方差分析中的主效应显著，可以用 Scheffe 检验。Scheffe 检验可以用在各种事先检验和事后检验中，包括成对检验和非成对检验，还可以用于当各组被试不相等的情况。

Scheffe 检验使用的是 F 分布。当进行事后检验时，F 检验的公式是：

$$F=\frac{(\overline{X}_1-\overline{X}_2)^2}{MS_{err}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

其中 $\overline{X}_1$、$\overline{X}_2$ 指参加比较的平均数，MS_{err} 是全方差分析中的误差项，n 是每组的被试数。

在 Scheffe 检验中，两个重要的参数同样是参加比较的组数和 MS_{err} 的自由度。临界值的查表的方法如下：

$$F_c=(K-1)[F_a(K-1,\ N-K)]$$

其中 K 是参加比较的组数，N 是总样本数，F_a 是根据自由度查表得到 F 值，F_c 是校正后的 F 值。

Scheffe 检验还可以用于事先比较检验，F 检验的公式是：

$$F=\frac{\left(\sum_{j=1}^{p}C_j\overline{X}_j\right)^2}{MS_{err}\sum_{j=1}^{p}\frac{C_j^2}{n_j}}$$

其中 $\overline{X}_j$ 是参加比较的平均数，C_j 是比较系数，n_j 是各组被试的数量，MS_{err} 是全方差分析中得到的误差项。

临界值的查表方法如下：

$$F_c=F_a(1,\ N-K)$$

其中 K、N、F_a 和 F_c 的含义与事后检验时相同。

我们举例来说明 Scheffe 检验的方法。假如我们的实验中有三组平均数，总样本数是 83 个数据。当进行一个事后比较检验（注意：该事后比较包括对三组平均数之间所有可能的两两比较，$K=3$）时，查 F 表可知：

当 $\alpha=0.05$ 时，$F(2,\ 80)=3.11$；

当 $\alpha=0.01$ 时，$F(2,\ 80)=4.88$。

根据临界值查表的公式 $F_c=(K-1)[F_a(K-1,\ N-K)]$，其临界值应当是：

当 $\alpha=0.05$ 时，$F(2,\ 80)=2\times3.11=6.22$；

当 $\alpha=0.01$ 时，$F(2,\ 80)=2\times4.88=9.76$。

如果我们对同样的这组数据进行一个事先比较检验，根据临界值查表

公式 $F_c = F_a(1, N-K)$，可知临界值应当是：

当 $\alpha=0.05$ 时，$F(1, 80)=3.96$；

当 $\alpha=0.01$ 时，$F(1, 80)=6.96$。

需要注意的是，在事先比较检验中使用 Scheffe 检验的前提是：当在一个含有多组平均数的实验中，研究者只对一个或少数几个成对或非成对比较进行事先检验。

可以看到，Scheffe 检验使用了非常严格的校正，而且随着参加比较的平均数的增加，临界值迅速提高。因此，与 Tukey-HSD 检验相比，Scheffe 检验更不容易拒绝虚无假说，或者说其检验力远远不如前者。与 ANOVA 类似，Scheffe 使用 F 取样分布，因此需要遵循 F 检验的前提。我们还可以看到，用 Scheffe 检验进行事先比较检验和事后比较检验的严格程度是不相同的。与事先比较相比，Scheffe 检验对事后比较进行了更加严格的校正，因为它假设事先比较的数量要少于事后比较。

（五）Bonferroni-Dunn 检验

Bonferroni-Dunn 检验的特点是在使用 t 检验时考虑控制实验错误率。在第十章已经提到，在一个实验中，随着独立比较的数目增加时，实验错误率迅速增加。

我们知道，

$$\alpha_{EW} = c \times \alpha_{PC}$$

可以将公式改写为：

$$\alpha_{PC} = \alpha_{EW} / c$$

从公式中可以看出，我们可以通过控制 α_{PC}，以便将 α_{EW} 限定在一个合适的水平，从而控制实验错误率。假如我们设 $\alpha_{EW}=0.05$，实验中有五个计划的比较，每一个 α_{PC} 应当设置为 $0.05/5=0.01$。这时实验中的 α_{PC} 和 α_{EW} 都被限定在我们预想的可以接受的Ⅰ型错误之内。

用 Bonferroni-Dunn 检验表可以查出其显著性。SPSS 软件中也会自动给出显著性结果。需要注意的是：当各组变异齐性时，选择使用合并变异（pooled variance，equal variances assumed）；当各组变异不齐性时，选择不要合并变异（equal variances not assumed）。

二、几种事后检验方法的比较

上面介绍的几种事后检验方法，通过不同的途径，对多重比较中的累积错误进行校正，以便将实验错误率控制在合适的水平。当参与比较的平均数数目增加时，不同的事后检验方法校正的程度有多大的差别？我们举例来说明。从表 11-5 中的举例中，可以看出在上述的几种事后检验方法中，当多重比较中的平均数数目达到 10 个时，LSD、Tukey-HSD 和 Scheffe 检验对所有的成对比较的检验标准是恒定的。LSD 检验使用的是最小临界值（1.98），相当于没有校正。Tukey-HSD 检验使用的是最大临界值（3.22），即使用了严格的校正。而 Newman-Keuls 检验考虑了比较是在平均数的邻近组之间（$r=2$）进行的还是在不邻近组之间（$r>2$）进行的。随着 r 的增加，Newman-Keuls 检验的临界值从 1.98 增加到 3.22，对不邻近组的平均数比较使用了更严格的校正。Tukey-HSD 检验相当于 Newman-Keuls 检验中当平均数比较是在最不邻近的组之间进行的。Scheffe 检验则设置了最严格的校正。

表 11-5　几种事后检验方法的临界值比较

r	2	3	4	5	6	7	8	9	10
LSD	1.98	1.98	1.98	1.98	1.98	1.98	1.98	1.98	1.98
Newman-Keuls	1.98	2.38	2.60	2.77	2.90	3.00	3.08	3.16	3.22
Tukey-HSD	3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.22
Scheffe	4.20	4.20	4.20	4.20	4.20	4.20	4.20	4.20	4.20

三、选择多重比较的检验方法

在研究中如何选择使用多重比较的检验方法？可以说，没有哪一种多重比较检验方法是最优的，或者说没有一种方法在任何情况下都比其他检验方法更加有效。选择一种合适的检验方法，主要与研究者要控制的 I 型错误率有关，同时也与研究者选择成对比较或复杂对比有关。

一般来说，选择多重比较的检验方法需要考虑以下的因素。

第一，研究中平均数的数量。如果只有两个平均数（$p=2$），我们可以方便地使用 t 检验或单因素方差分析的 F 检验，不涉及多重比较的问题。当平均数多于两个（$p>2$）时，则需要考虑进行多重比较检验。

第二，关于平均数比较的假说是事先的还是事后的。如果要检验一个关于多个平均数之间比较的事先计划的假说，可以使用对比的方法。另外还需要考虑检验的是一组正交的比较还是一组非正交的比较，进一步选用相应的统计方法。事先比较的检验主要是用 t 检验的方法。

第三，如果在全方差分析显著后进行一个关于多个平均数之间的事后检验，Tukey-HSD 检验、Newman-Keuls 检验和 Scheffe 检验都是可选用的。相比之下，当研究者只关心成对比较时，Tukey-HSD 检验和 Newman-Keuls 检验是较好的选择，因为 Scheffe 检验更加严格，不容易达到显著。但 Scheffe 检验的优点是，它不仅可以用于成对比较的检验，也可以用于非成对比较的检验，或者用于成对与非成对比较检验混合的情况。

另外，随着一个研究中比较数量的增加，Ⅰ型错误增加。实验错误率和每个比较错误率的关系是：

$$\alpha_{EW}=c\times\alpha_{PC}$$

从关系公式可以看出，如果我们想保持实验的总错误率 α 在某个水平，可行的方法有两个方面：一个是减少比较的数量 c，另一个是减小每个比较的错误率（α_{PC}）。这就是各种多重比较校正技术所试图做的。

我们需要记住的是，两类错误对经验科学发展的影响是不同的。减小Ⅰ型错误的结果会导致Ⅱ型错误的增加。增加Ⅱ型错误，使研究者更可能看不到实验数据中真实存在的差异。但是如果我们减少Ⅱ型错误，增加Ⅰ型错误，我们可能增加了发现某些小的、研究者感兴趣的“真”效应的机会。我们需要对两类错误对科学实验的影响各是什么的问题有更深入的认识，以便在两类错误之间进行平衡。研究者通常比较重视Ⅰ型错误，希望设置更严格的拒绝区域，以减少虚报的错误。在科学研究的过程中，Ⅰ型错误通常是可以在重复实验中得到纠正的。在一个活跃的研究领域，当我们重复了他人的实验并得到同样的实验结果时，我们可以非常自信“拒绝假说”的决定是正确的。因为在这种独立的验证中，我们的结果是“错误”的概率是很小的。另外，重复实验经常不是完全精确的重复，而是会集中操作一些关键的特征，随着实验的重复，我们就会认识到原实验中的不正确的结果。或者说，在原实验中对虚无假说的不正确的拒绝经常会在

重复实验中得到纠正，因为如果后来的许多研究者的独立实验都无法重复原实验的结果，就会质疑原实验所报告的显著差异，从而纠正原实验中的Ⅰ型错误。

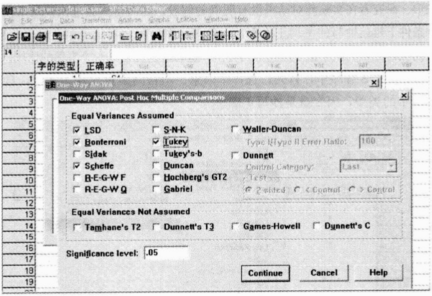
第二节　事后比较的SPSS常用计算方法

在一个自变量的水平数大于2的研究中，方差分析主效应显著之后，采用多重比较检验确定差异发生在哪两个水平之间时，可能增大Ⅰ型错误。因此，出现了许多对Ⅰ型错误累积的校正方法，方法多达18种以上。SPSS软件中常用的方法有LSD检验、Sheffe检验、Bonferroni-Dunn检验、Tukey-HSD检验、Newman-Keuls检验等。

LSD检验等同于进行多个成对t检验，它没有控制犯Ⅰ型错误的概率。Bonferroni-Dunn检验在比较的数目不太多的情况下是很好的，在比较数目较多的时候会增大Ⅱ型错误。除了Bonferroni-Dunn检验，Scheffe检验则是最保守的检验之一，它控制整个实验中的比较犯Ⅰ型错误的概率始终在5%内。Scheffe检验既可以作成对比较，也可以作非成对的对比，但是在SPSS中只提供了成对比较的方法。下面我们举例介绍几种事后的多重比较的SPSS操作。

一、事后的多重比较的SPSS操作

在SPSS One-Way提供的18种事后比较（unplanned comparisons/posteriori tests/post hoc）中，主要分两大类：方差齐性满足时的14种和方差齐性不满足时的4种比较方法。它们大体提供了两种格式的信息。有些只是区分同质组，如Student Newman-Keuls（S-N-K）检验、Tukey's-b检验；有些只作多重比较，如LSD检验、Bonferroni-Dunn检验；有些是两种都提供，如Tukey-HSD检验、Scheffe检验。常用的有LSD检验、Sheffe检验、Bonferroni-Dunn检验、Tukey-HSD检验、S-N-K检验等。



事后比较采用的统计方法主要有三种：q 检验、t 检验和 F 检验。Tukey-HSD 检验、Newman-Keuls 检验、Duncan's New Multiple 检验等事后检验主要使用 q 检验。LSD 检验、Bonferroni-Dunn 检验、Dunnett 检验等主要使用 t 检验。Scheffe 检验主要使用 F 检验。

三种统计方法的计算公式如下：

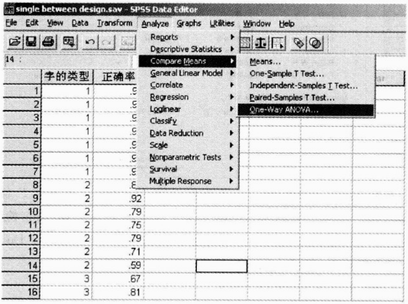
$$t=\frac{\overline{X_1}-\overline{X_2}}{\sqrt{MS_{err}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad q=\frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{MS_{err}}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad F=\frac{(\overline{X_1}-\overline{X_2})^2}{MS_{err}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

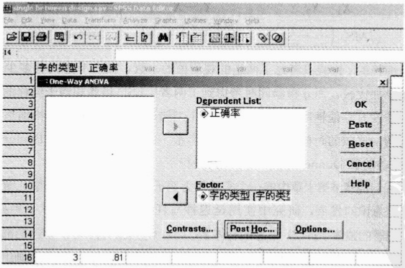
纯粹的 t 检验只考虑了两组的观测值的个数，没有考虑共有多少组平均数参加比较，而 q 检验同时考虑到了两者。LSD 检验相当于纯粹的 t 检验，并没有其他辅助校正，所以它犯 Ⅰ 型错误的概率会较大，特别是在作多组两两比较的时候，这个问题尤其严重。而其他两种检验，Bonferroni-Dunn 检验和 Dunnett 检验，则对其进行了一定的校正。

我们仍然举第十章中的一个实验作为例子。研究者要探讨小学儿童汉字正字法意识的发展，研究中使用延迟抄写任务。实验中设计了五种条件：(1) 熟悉的独体字；(2) 熟悉的合体字；(3) 不熟悉的合体字，但字的部件是熟悉的；(4) 不熟悉的合体字，字的部件也是不熟悉的；(5) 随机笔画组成的

图形。研究者在实验前对结果没有明确的假设，因而希望通过比较儿童在各种条件下延迟抄写的平均正确率，检验平均数两两之间是否存在显著差异，从而探讨儿童的汉字正字法意识的特点。这时要进行的是一个事后的比较。

在 SPSS 程序中，首先，选择单因素、项目间多重比较。其中，自变量是字的类型，因变量是抄写正确率。





对儿童的汉字正字法意识实验的例子，使用 Tukey-HSD 检验进行成对比较的输出结果如下（见图 11-1）。

Output3 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Post Hoc Tests

Multiple Comparisons

Dependent Variable: 正确率

	(I) 字的类型	(J) 字的类型	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
Tukey HSD	1.00	2.00	.20286*	.06400	.027	.0172	.3885
		3.00	.23714*	.06400	.007	.0515	.4228
		4.00	.61429*	.06400	.000	.4286	.7999
		5.00	.51857*	.06400	.000	.3329	.7042
	2.00	1.00	-.20286*	.06400	.027	-.3885	-.0172
		3.00	.03429	.06400	.983	-.1514	.2199
		4.00	.41143*	.06400	.000	.2258	.5971
		5.00	.31571*	.06400	.000	.1301	.5014
	3.00	1.00	-.23714*	.06400	.007	-.4228	-.0515
		2.00	-.03429	.06400	.983	-.2199	.1514
		4.00	.37714*	.06400	.000	.1915	.5628
		5.00	.28143*	.06400	.001	.0958	.4671
	4.00	1.00	-.61429*	.06400	.000	-.7999	-.4286
		2.00	-.41143*	.06400	.000	-.5971	-.2258
		3.00	-.37714*	.06400	.000	-.5628	-.1915
		5.00	-.09571	.06400	.573	-.2814	.0899
	5.00	1.00	-.51857*	.06400	.000	-.7042	-.3329
		2.00	-.31571*	.06400	.000	-.5014	-.1301
		3.00	-.28143*	.06400	.001	-.4671	-.0958
		4.00	.09571	.06400	.573	-.0899	.2814

图 11-1 Tukey-HSD 成对比较的输出结果

从图 11-1 的结果可以看出，Tukey-HSD 检验中共进行了 10 个成对比较，除了两个比较，即字的类型 2（熟悉的合体字）和字的类型 3（不熟悉的合体字，但字的部件是熟悉的），以及字的类型 4（不熟悉的合体字，字的部件也是不熟悉的）和字的类型 5（随机笔画组成的图形）之间差异不显著外，其他的比较都是显著的。

图 11-1 和表 11-6 中两种形式的结果是等价的，在 Homogeneous Subsets 里，它把组间差异在 0.05 水平上不显著的组归为一个分组（subset），并将差异最大的两组的差异显著性标在相应的小组里。

表 11-6 Tukey-HSD 检验和 Scheffe 检验的输出结果

Homogeneous Subsets

正确率

	字的类型	N	Subset for alpha = .05		
			1	2	3
Tukey HSD[a]	4	7	.3629		
	5	7	.4586		
	3	7		.7400	
	2	7		.7743	
	1	7			.9771
	Sig.		.573	.983	1.000
Scheffe[a]	4	7	.3629		
	5	7	.4586		
	3	7		.7400	
	2	7		.7743	.7743
	1	7			.9771
	Sig.		.694	.990	.063

Means for groups in homogeneous subsets are displayed.

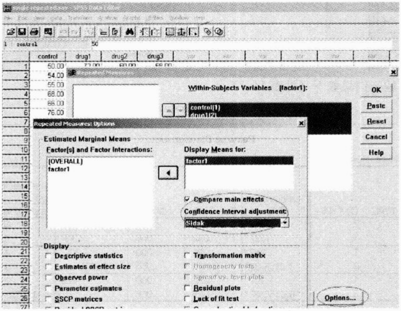
a. Uses Harmonic Mean Sample Size = 7.000.

表 11-6 中同时显示了 Tukey-HSD 检验和 Scheffe 检验的结果。Tukey-HSD 检验中，除字的类型 2（熟悉的合体字）和字的类型 3（不熟悉的合体字，但字的部件是熟悉的）之间，字的类型 4（不熟悉的合体字，字的部件也是不熟悉的）和字的类型 5（随机笔画组成的图形）之间差异不显著外，其他比较都是显著的。在 Scheffe 检验中，则有三组比较不显著：字的类型 2（熟悉的合体字）和字的类型 3（不熟悉的合体字，但字的部件是熟悉的）之间，字的类型 4（不熟悉的合体字，字的部件也是不熟悉的）和字的类型 5（随机笔画组成的图形）之间，以及字的类型 1（熟悉的独体字）和字的类型 2（熟悉的合体字）之间差异不显著，其他比较都是显著的。

在 SPSS 中，被试间（项目间）因素的多重比较都可以通过选项“post-hoc”来实现，只是在具体选择哪种方法上要根据自己的实验目的而定，究竟实验设计是需要尽量控制 I 型错误还是需要尽量减少 II 型错误。从上面的计算中，我们可以再一次看到几个多重比较之间的关系：相对于 Tukey-HSD 检验，Scheffe 检验使用了更严格的标准。读者自己还可以尝试其他方法，结果会发现：LSD 检验是最宽松的检验，它对因多个比

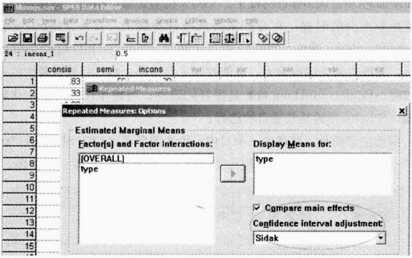
较而增加的I型错误不作任何校正。Scheffe 检验是最严格的，特别是当多个比较是非正交的时候，它会严重降低统计检验力。有人（Petrinovich & Hardyck，1969）建议：当进行成对比较时，Tukey-HSD 检验是较好的方法，它既有较好的检验力，也能控制Ⅰ型错误；当进行数目较少的正交比较时，Bonferroni-Dunn 检验是很好的选择；当想作任意的多个比较的时候，Scheffe 检验是较适合的选择。

需要注意的是，被试内（项目内）因素的多重比较在 SPSS 的“post-hoc”里面是没有的。它把这项功能放到了“options”里面，见下图。



我们还是以第十章中的实验数据为例看事后的比较在 SPSS 中如何实现。在关于儿童的汉字读音中声旁规则性发展的实验中，我们考察的因素，即字的类型，有三个水平：规则字、半规则字和不规则字。我们事先没有特定的假设，希望通过穷尽所有的比较，探讨各种字之间是否有差异。这是一组事后的比较。

在 SPSS 程序中，我们进行事后的多重比较，也可以使用前面提到的 Compare main effects 来做（如下图），我们选用 Sidak。



输出结果如下（见表 11-7）。

表 11-7 事后检验的输出结果

Pairwise Comparisons

Measure: MEASURE_1

(I) TYPE	(J) TYPE	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a] Lower Bound	Upper Bound
1	2	.247*	.035	.000	.161	.332
	3	.327*	.032	.000	.248	.406
2	1	-.247*	.035	.000	-.332	-.161
	3	8.015E-02*	.025	.006	1.975E-02	.141
3	1	-.327*	.032	.000	-.406	-.248
	2	-8.015E-02*	.025	.006	-.141	-1.975E-02

Based on estimated marginal means

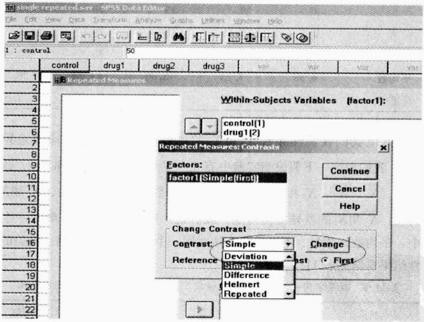
*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Sidak.

在事后比较中，我们只能进行两两成对比较，共进行了三个比较。可以看到：规则字和半规则字之间的差异是显著的（$p=0.000$）；规则字（1）和不规则字（3）之间的差异是显著的（$p=0.000$）；半规则字（2）和不规则字之间的差异也是显著的（$p=0.006$）。如果与第十章中的事先比较的统计检验显著性相比，事先正交对比中有两个比较的结果是显著的，即规则字和不规则字之间的差异是显著的（$p=0.000$），半规则字

和不规则字之间的差异是显著的（$p=0.002$）。可以看出，在这个实验中，事后比较检验比事先检验多进行了一个比较，然而，两种比较方法得到的结果是类似的。

除了单因素组间设计，其他的实验设计，比如单因素重复测量、多因素实验设计等，它们的比较在 SPSS 里也提供了一些选项，但是相对不灵活。用得比较多的是其中的 Simple，它只能作两两比较，提供了两种模式。一种是后面的各种条件都与第一种条件进行比较，即 Reference Category 选 First；另一种是前面各种条件都与最后一种条件进行比较，即 Reference Category 选 Last。这样的比较在包含几个实验组与一个控制组的实验设计中是很有用的，它提供了几个实验组与同一个控制组分别比较的方法。其他几种比较方法在此不作叙述。



二、各种事后比较检验方法的异同

前面提到，各种 post-hoc 在处理累积错误上的标准是不一样的，而累积错误的问题随着实验中平均数的数目的增加而变得越来越严重。下面我们拟采用不同组数比较的 post-hoc 来看看各种检验的结果差异情况。

在一个单因素完全随机实验中，有五个处理水平，即 A1，A2，A3，A4，A5，各处理条件下都是八个被试，平均数和标准差的情况如表 11-8。我们举例来比较一下，当一个实验中平均数的数目不同、使用不同的检验方法时结果的差异。

表 11-8 六种处理条件下被试的原始数据

处理条件	A1	A2	A3	A4	A5
1	3	8	10	15	18
2	5	7	14	18	19
3	4	9	14	16	20
4	8	6	12	17	20
5	6	7	13	18	21
6	4	8	15	19	19
7	5	5	12	20	22
8	4	7	14	15	23
平均数	4.88	7.13	13.00	17.25	20.25
标准差	1.55	1.25	1.60	1.83	1.67

在这里我们仅关注条件$\overline{X_1}$和$\overline{X_2}$的比较（$\overline{X_1}$和$\overline{X_2}$的平均数差异为 2.25）。我们考察一下假设当实验中有三个平均数（$\overline{X_1}$、$\overline{X_2}$和$\overline{X_3}$）、四个平均数（$\overline{X_1}$、$\overline{X_2}$、$\overline{X_3}$和$\overline{X_4}$）以及五个平均数（$\overline{X_1}$、$\overline{X_2}$、$\overline{X_3}$、$\overline{X_4}$和$\overline{X_5}$）的时候，$\overline{X_1}$和$\overline{X_2}$的差异检验情况。或者说，我们希望读者了解在一个实验中，平均数的数目、不同检验方法对同一个比较结果的影响。检验的显著性结果总结见表 11-9。

表 11-9 数量不同平均数参加比较情况下$\overline{X_1}$和$\overline{X_2}$差异的事后检验显著性

	三个平均数	四个平均数	五个平均数
LSD	0.006**	0.008**	0.008**
Tukey-HSD	0.016*	0.037*	0.056
Sidak	0.018*	0.046*	0.075
Bonferroni-Dunn	0.018*	0.047*	0.077
Scheffe	0.021*	0.063	0.117

从表 11-9 中可以看到，无论在三个、四个还是五个平均数的情况下，对$\overline{X_1}$和$\overline{X_2}$差异的 LSD 检验显著性基本是不变的。可以明显看到，运用 LSD 检验在对两个平均数的差异进行检验时，是不考虑其他的平均数的情况的。或者说，在一个实验中，无论有几个平均数之间的比较，对某一个比较来说，LSD 检验的结果是不变的。然而，在其他四种检验中，可以明显看到随着实验中平均数比较数目的增加，同一个比较的检验显著性下降。其中 Scheffe 检验是最严格的，当实验中存在四个平均数或五个平均数时，$\overline{X_1}$和$\overline{X_2}$的差异都变得不显著了（$p=0.063$，$p=0.117$）。Bonferroni-Dunn 检验和 Sidak 检验其次，当实验中存在五个平均数时，$\overline{X_1}$和$\overline{X_2}$的差异变得不显著（$p=0.077$，$p=0.075$）。相对来说，Tukey-HSD 检验比较容易显著一些，当实验中存在五个平均数时，$\overline{X_1}$和$\overline{X_2}$的差异变得边缘显著（$p=0.056$）。而 LSD 检验是最容易显著的，当实验中存在五个平均数时，$\overline{X_1}$和$\overline{X_2}$的差异仍然是非常显著的（$p=0.008$）。从这个表中，我们可以清楚地看到，为了控制实验Ⅰ型错误率，各种检验方法对比较进行了不同程度的校正。

三、事先对比和事后比较的优缺点

从以上的例子和讨论中，我们可以看到，事先对比的优越性表现在以下两个方面。(1) 事先对比是对某种理论、假说的直接检验，往往是在理论的指导下进行的。使用事先对比促使实验者在实验设计阶段对实验结果、预期有全面深入的思考，通过设置特定虚无假说直接提问感兴趣的问题，得到的检验结果能更好地从理论上进行解释。(2) 由于事先对比用最少的比较获得了对研究者感兴趣的理论问题的检验，因此累积错误的问题相对比较小。由于在对Ⅰ型错误进行校正的同时也会降低检验力，因而与事后比较相比，事先对比增加了检验力，容易发现平均数之间实际存在的差异。鉴于以上特点，在可能的情况下，研究者往往采用事先对比，并根据关心问题的特点选用正交的或非正交的对比。

然而，在很多研究中，研究者在事前并没有明确的理论假设，而是需要根据统计检验的结果模式发现存在的差异或发展理论，这时事后比较也

是非常有效的。事后比较通过穷尽所有的成对比较，发现和解释差异，从而导致新结果的发现。在事后比较技术中，统计学家已经对Ⅰ型错误进行了校正，并且可以方便地使用现成的SPSS软件进行计算，因此事后比较也是研究者最常使用的方法。

本章主要观点

• 对多重比较中的累积错误进行校正，有三条解决问题的途径：对统计结果不进行任何校正；对统计结果进行非常严格的校正；处于两者中间的途径则是校正的程度依赖于被比较的平均数的数量。

• LSD检验的特点是对全方差分析显著后所作的比较不作任何校正。Tukey-HSD检验的目的是对所有可能的比较控制实验错误率，它对所有的比较使用相同的严格的校正。Newman-Keuls检验的校正使得每个比较，甚至极端组的比较，都被调整在 α 水平。与Tukey-HSD检验相比，Newman-Keuls检验更容易拒绝虚无假说。Scheffe检验使用了非常严格的校正，而且随着参加比较的平均数的增加，临界值迅速提高。

• 选择一种合适多重比较的检验方法，与研究者要控制的Ⅰ型错误率有关，同时也与研究者选择成对比较或复杂对比有关。

• 选择多重比较的检验方法需要考虑的因素包括：研究中平均数的数量，研究中关于平均数比较的假说是事先的还是事后的，以及研究检验成对还是非成对比较。当研究者只关心成对比较时，Tukey-HSD检验和Newman-Keuls检验是较好的选择。Scheffe检验可以用于成对比较、非成对比较、成对与非成对比较混合的情况。

• 从实验错误率和每个比较错误率的关系公式可以看出，要保持实验的总错误率 α 在某个水平，可以通过两个途径：一个是减少比较的数量；另一个是减小每个比较的错误率。这就是各种多重比较校正技术所试图做的。

思考题

1. 什么是事后比较？在什么情况下使用事后比较？为什么事后比较

的累积错误问题更加严重？

2. 事后比较有哪些常用的统计检验方法？这些方法是通过什么途径进行累积错误校正的？

3. LSD检验的特点和过程是什么？

4. Tukey-HSD 检验和 Newman-Keuls 检验的特点和过程分别是什么？

5. Scheffe 检验的特点和过程是什么？

6. Bonferroni-Dunn 检验的特点和过程是什么？

7. 各种事后多重比较方法的优缺点各是什么？如何选择合适的多重比较方法？

第十二章
复杂的实验设计和数据分析

前面的章节已经介绍了如何通过实验设计来减少实验误差的方法。然而在心理、教育、社会研究中，研究者经常会遇到一些更复杂的情况，导致我们使用普通的实验设计和方差分析技术无法分离出一些在实验中产生影响的无关变异，需要使用一些更复杂的实验设计和统计方法。本章介绍的嵌套实验设计、协方差分析将对解决这样的问题有所帮助。

第一节　嵌套实验设计

在本书的前几章中，我们已经了解了各种实验设计的一些重要原则。例如，我们知道，要得到真实的处理效应，在非重复测量或被试间实验设计中，重要的前提是能够随机分配被试，以保证在施测前接受不同实验处理的各组被试的差异可以忽略不计。在重复测量或被试内实验设计中，重要的前提是当一个被试接受多次实验处理时，不同自变量水平的先后施测对同一被试不产生长期影响效应。但是在心理、教育、社会研究中，研究者经常会遇到这两个条件都不能满足的情况。一些研究中，实验处理的实施具有学习、记忆效应，因此不适宜进行重复测量实验设计。例如，在比较两种教学方法的研究中，实施第一种教学方法会对第二种教学方法的效果产生影响，因此比较两种教学方法的效应不可能通过在一个被试身上施测而得到，只能进行被试间实验设计。然而，许多心理、教育、社会研究又常常是在一些固定的社会团体中进行的。如教学实验经常是在特定的班级中进行的，药物实验经常是在特定的医院、病房中进行的。在实践中，班级、车间、病房都是固定的社会团体，将这样团体中的成员随机分配给各个实验处理有时是不可能的。因此，使用一般的完全随机实验设计也是

有困难的。在这样的情况下，嵌套实验设计（nested design）是解决上述问题的途径之一。还有一些情况下，实验设计中一些无关因素可能影响因变量的测量，但它们不是研究者感兴趣的。嵌套实验设计对解决这种问题也会有所帮助。

嵌套实验设计与前面几章介绍的实验设计不同。在前面介绍的多因素实验设计中，因素的水平之间是交叉的，即一个因素的每个水平与另一个因素的所有水平相结合。例如，在一个 2×3 两因素完全随机实验设计中，有 A1B1、A1B2、A1B3、A2B1、A2B2、A2B3 共六个处理水平的结合。在这样的实验设计中，我们可以观察、计算两个或多个因素的交互作用，所以也可以称这类实验设计为交叉实验设计。然而，在嵌套实验设计中，一个因素的每个水平仅出现在另一个因素的某个水平上。例如，在一个两因素嵌套实验设计中，可能有的处理水平包括 A1B1、A1B2、A1B3、A2B4、A2B5、A2B6。虽然也是六个处理水平的结合，但我们可以明显看出这六个处理水平的结合与交叉实验设计中的六个处理水平的结合是不一样的。在嵌套实验设计中，事先假设两个因素之间不存在交互作用。我们通过一个两因素完全随机设计和一个两因素嵌套设计分配被试的图解来看它们的差别。

表 12-1　两因素完全随机设计和两因素嵌套设计分配被试的图解

两因素完全随机设计被试分配方案			两因素嵌套实验设计被试分配方案		
	A1	A2		A1	A2
B1	√	√	B1	√	
			B2	√	
B2	√	√	B3	√	
			B4		√
B3	√	√	B5		√
			B6		√

注：√表示在该处理水平的结合中分配了被试。

从图解中可以看到，在一个两因素完全随机设计中，A 因素的两个水平与 B 因素的三个水平相结合，共有六种处理水平的结合，被试被随机分配给这些实验处理的结合。然而，在一个两因素嵌套设计中，B 因素的 B1、B2、B3 三个水平仅出现在 A1 水平，而 B 因素的另外三个水平 B4、B5、B6 仅出现在 A2 水平。这时，我们说 B 因素是被嵌套在 A 因素

中的。因此，嵌套设计是指在因素实验设计中，至少一个因素的水平是被局限在另一个因素的水平中的。例如，如果 B 因素的每个水平仅出现在 A 因素的一个水平中，B 就是嵌套于 A 的，可写作 B(A)。

在本章中，我们将介绍嵌套实验设计经常适用的两种情况。第一种是被试组在实验处理条件中的嵌套，在这种情况下，被嵌套的因素通常指实验中使用的固定团体，嵌套设计的目的主要是分离出无关变量——团体效应，以便精确地估价实验处理的效应。第二种是控制因素或无关因素在实验处理条件中的嵌套。在这种情况下，被嵌套的因素通常指实验中的无关变量。当实验中的无关变量可能影响因变量的测量，因而与实验处理效应相混淆时，嵌套设计可以帮助分离出无关变量的效应。

一、被试组在处理条件中的嵌套

在有些实验中，将一些固定团体的被试组分配给不同的处理条件是更方便和更理想的选择。行为研究中，研究者经常需要从一个更大的总体中抽取成组的被试，如从学校中抽取整班级的学生，从城市的不同医院中抽取整病房的病人，从不同的工厂中抽取整车间的工人。例如，要探讨教学方法的效应，更好的实施方案是分配不同班级的学生接受不同的教学方法。这时，对被试进行固定团体的施测比进行随机分配的小组施测更方便。在这种情况下，如果实验处理是两种或多种教学方法，将两个或多个固定团体被试组分配给每一个实验处理条件，就叫做被试组在实验处理条件中的嵌套设计。

有时，也许不一定是需要团体测试被试，而只是因为来自一个团体的被试接受了某一种特殊的实验处理，而来自其他团体的被试接受了其他的实验处理。例如，来自一个医院的病人接受了某一种药物的治疗，而来自其他医院的病人接受了其他药物的治疗。也就是说，来自每个医院的病人仅接受了一种药物的处理，这时，医院是被嵌套在药物治疗的处理条件中的。

（一）两因素完全随机嵌套实验设计

最简单的嵌套设计中含一个自变量和一个团体变量，研究条件通常有以下的特点。

(1) 研究是在固定的团体中进行的。由于不能随机分配被试给各个实验处理，研究中有两个可能影响因变量的因素，一个是自变量，另一个是无关变量——固定的团体。假设两个因素之间没有交互作用。

(2) 研究中一个因素的水平是嵌套在另一个因素的水平之中的。如果自变量有两个水平，无关变量有六个水平，实验中共有六个处理的结合。

实验设计的基本方法是，将被嵌套因素（团体）的各个水平随机分配给另一个因素（实验处理）的水平中。例如，在表 12-2 中，把 B 因素的六个水平随机分配给 A 因素的两个水平，即随机分配三个团体接受 A1 水平的处理，另外的三个团体接受 A2 水平的处理。这时，每个团体中的被试不再进行随机分配。

例如，在一个教学实验中，要考察三种教学方法对学生学习效果的影响。有六个班的 144 名学生参加了实验。一种实施的方法是在自然班级中对学生进行教学方法的实验，每个班 24 名学生接受一种教学方法。另一种做法是将六个班中每个班的 24 名学生随机分成三组，对每组学生进行一种教学方法实验。让我们来比较两种方法：

表 12-2　嵌套设计的被试分配

		教学方法(A)	1	2	3
班级(B)	1		24		
	2		24		
	3			24	
	4			24	
	5				24
	6				24

表 12-3　完全随机设计的被试分配

		教学方法(A)	1	2	3
班级(B)	1		8	8	8
	2		8	8	8
	3		8	8	8
	4		8	8	8
	5		8	8	8
	6		8	8	8

可以看出，前一种方法在实施上更方便，做法也更加接近自然。在一些研究情境下，尤其是教育、社会等研究中，嵌套实验设计是唯一的选择。从表中可以看出，在一个完全随机嵌套设计中，只有部分因素水平的结合被测量了，因此不能计算交互作用。如果团体的数量是 q，每个团体的被试数是 n，且各团体人数是平衡的，则实验需要的总被试量是 $N=qn$。

两因素完全随机嵌套设计可检验的假说是：

(1) $H_0: \alpha_j=0$

即 A 因素的处理效应等于 0；

(2) $H_0: \sigma_b^2=0$

即团体效应等于 0。

它的实验设计模型是：

$$Y_{ij}=\mu+\alpha_j+\beta_{k(j)}+\varepsilon_{i(jk)}$$

$$(i=1, 2, \cdots, n; j=1, 2, \cdots, p; k=1, 2, \cdots, q)$$

其中 μ 为总体平均数或真值，α_j 为 A 因素的水平的处理效应，$\beta_{k(j)}$ 为嵌套在 A 因素水平 j 中的 B 因素第 k 个水平的效应，$\varepsilon_{i(jk)}$ 为误差变异。

从实验设计模型中可以看出，交互作用在这里是不出现的。$\beta_{k(j)}$ 或团体变异是无关变异。团体变量被包含在实验设计中，因为研究者假设它可能影响因变量的观测值，但它本身不是研究者感兴趣的变量。

（二）三因素完全随机嵌套实验设计

更复杂的嵌套实验设计涉及两级的固定团体。例如，被试取样来自不同的学校中不同的班级。三因素完全随机嵌套设计适用于下列研究条件。

(1) 研究是在固定的团体中进行的，由于上下两级团体都可能影响因变量的观测值，因此实验中实际上含一个自变量和两个无关变量。假设自变量和无关变量之间没有交互作用。

(2) 研究中的一个因素（高一级团体，如学校）是嵌套在另一个因素（实验处理）的水平中的，第三个因素（低一级团体，如班级）是既嵌套在高一级团体（如学校）又嵌套在实验处理中的。

例如，在一个激励方式的实验中，要考察新的激励方式与传统激励方式对职工工作效率的影响。有四个工厂中八个车间的 160 名职工参加了实验。一种实施的方法是在自然的工厂、车间中对职工进行激励方式的实

验，每个车间接受一种激励方式（新的激励方式或传统激励方式）。另一种做法是将每个车间中的 20 名职工随机分成两组，对每组进行一种激励方式的实验。

嵌套实验的基本方法是首先将高一级团体（如工厂）的四个水平随机分配给实验处理因素，即随机分配四个工厂接受不同的激励方式的处理。这时低一级团体的八个水平随之被分配给了不同的实验处理，即两个工厂的四个车间接受新的激励方式的处理，另两个工厂的四个车间接受传统激励方式的处理。每个低一级团体内的被试不再进行随机分配。

让我们来比较两种方法：

表 12-4 嵌套设计的被试分配

		激励方式(A)	1	2
工厂(B)	车间(C)			
1	1		20	
1	2		20	
2	3		20	
2	4		20	
3	5			20
3	6			20
4	7			20
4	8			20

表 12-5 完全随机设计的被试分配

		激励方式(A)	1	2
工厂(B)	车间(C)			
1	1		10	10
1	2		10	10
2	3		10	10
2	4		10	10
3	1		10	10
3	2		10	10
4	3		10	10
4	4		10	10

可以看出，在实际的实验实施中，第一种研究实施方法或嵌套设计方法是更可行的。

三因素完全随机嵌套设计可检验的假说是：

(1) $H_0: \alpha_j = 0$

即A因素的处理效应等于0；

(2) $H_0: \sigma_b^2 = 0$

即高一级团体效应等于0；

(3) $H_0: \sigma_c^2 = 0$

即低一级团体效应等于0。

它的实验设计模型是：

$$Y_{ij} = \mu + \alpha_j + \beta_{k(j)} + r_{l(jk)} + \varepsilon_{i(jkl)}$$

$(i=1, 2, \cdots, n;\ j=1, 2, \cdots, p;\ k=1, 2, \cdots, q;\ l=1, 2, \cdots, r)$

其中 μ 为总体平均数或真值，α_j 为A因素水平的处理效应，$\beta_{k(j)}$ 为嵌套在A因素 j 水平中的B因素第 k 个水平的效应或高一级团体的效应，$r_{l(jk)}$ 为嵌套在A因素 j 水平和B因素 k 水平中的C因素第 l 个水平的效应或低一级团体的效应，$\varepsilon_{i(jkl)}$ 为误差变异。

从实验设计模型中可以看出，交互作用在三因素嵌套实验设计中是不出现的，这也是使用嵌套设计的一个前提条件。

二、无关因素在实验处理条件中的嵌套

一般嵌套实验设计多用于固定被试团体在实验处理条件中的嵌套，有时嵌套实验设计还可以用于分离实验中的其他无关因素，因此也有其他无关变量在实验处理条件中的嵌套。这样的设计经常出现在下列的情况，实验中有些无关变量不是研究者感兴趣的，但它们可能影响因变量的测量，与实验处理效应混淆。嵌套设计也可以帮助分离出这样的无关变量的效应。例如，在一个实验中，研究者要比较学生对说明文和叙述文的阅读理解，为了增加实验的可靠性和推广力，研究者可能在每种文章类型中使用多篇文章，而不是一篇文章。使用多篇文章带来的一个问题是，这些文章在其他特征上的差异，如熟悉性，可能也会影响学生的阅读理解，但是研究者并不对文章中哪篇更容易或更难阅读感兴趣，因此文章在其他特征上

的差异只是一个控制因素或无关因素，可以将它们嵌套在文章类型中。这种嵌套设计的方式如下：

表 12-6　嵌套设计的文章分配

	文章类型	说明文	叙述文
文章		1	
		2	
		3	
		4	
			5
			6
			7
			8

（一）两因素完全随机嵌套实验设计

我们来看一个嵌套实验设计的例子。研究者要考察练习强度对于学习困难儿童学习生字的影响。实验中，研究者单独教孩子学习不熟悉的汉字。不同的实验组中儿童的原有水平相当，按照不同的训练强度学习。因变量是学生最终掌握的生字的数量。A 因素是学习强度，有高度(A1)、中度(A2)、轻度(A3) 三个水平。由于训练场地的限制，只能分三天的上午、下午和晚上三个时间段进行。但是，不同的训练时间是否也会影响处理效应？由于训练时间是一个研究者不感兴趣的无关变量，为了分离出不同训练时间的效应，研究者把时间段设为 B 因素。共 27 名儿童参加了实验，研究者将每个实验处理条件下参与训练的九名儿童随机分为三组，每组三人，每个组的儿童在同一时间进行训练，因此每一组内的被试接受的实验处理是完全相同的，在不同时间段接受不同的实验处理的儿童有 B1 到 B9 共九个水平。分配被试的例子如下。

表 12-7 嵌套设计的训练时间和被试分配

训练强度	A1	A2	A3
按时间段的被试分组			
	B1		
	B2		
	B3		
		B4	
		B5	
		B6	
			B7
			B8
			B9

实验实施后，在不同时间段接受不同强度训练的儿童组掌握生字数量的结果如下：

表 12-8 按时间段分组学习强度不同的儿童掌握生字的数量

被试分组	B1	B2	B3	B4	B5	B6	B7	B8	B9	Σ
学习强度	$n=3$									
A1	24	18	21							63
A2				15	14	16				45
A3							6	12	12	30
Σ										138

为了方便起见，我们只根据表 12-8 中学习强度不同的被试组掌握生字数量总和计算平方和：

$$SS_A = \frac{\sum_{j=1}^{p} A^2}{nq_{(j)}} - \frac{T^2}{npq_{(j)}}$$

$$= \frac{63^2 + 45^2 + 30^2}{3 \times 3} - \frac{138^2}{3 \times 3 \times 3}$$

$$= 766 - 705.33$$

$$= 60.67$$

其中 A 是每个学习强度条件下掌握生字数量的总和，T 是所有学习强度条件下掌握生字数量的总和，n 是每种学习强度下每个时间段的被试数量，p 是学习强度的水平数量，$q_{(j)}$ 是嵌套在每种学习强度下的时间段。

$$
\begin{aligned}
SS_{B(A)} &= \sum_{j=1}^{p}\frac{\sum_{k=1}^{q_{(j)}}(AB)^2}{n}-\frac{\sum_{j=1}^{p}A^2}{nq_{(j)}} \\
&= \frac{24^2+18^2+\cdots+12^2+12^2}{3}-\frac{63^2+45^2+30^2}{3\times 3} \\
&= 780.67-766 \\
&= 14.67
\end{aligned}
$$

其中 AB 是每个被试组在某个学习强度条件下掌握生字数量的总和。$SS_{B(A)}$ 是总变异中由嵌套在 A 因素中的 B 因素引起的变异，在嵌套设计中是一种无关变异。

$$
\begin{aligned}
SS_{总变异} &= \sum_{j=1}^{p}\sum_{k=1}^{q_{(j)}}\sum_{i=1}^{n}(ABS)^2-\frac{T^2}{npq_{(j)}} \\
&= 808.70-\frac{138^2}{3\times 3\times 3} \\
&= 808.70-705.33 \\
&= 103.37
\end{aligned}
$$

其中 ABS 是每个被试在某个学习强度条件下掌握生字数量的原始分数。已知被试的原始分数的总平方和是 808.70。

$$
\begin{aligned}
SS_E &= SS_{总变异}-SS_A-SS_{B(A)} \\
&= 103.37-60.67-14.67 \\
&= 28.03
\end{aligned}
$$

表 12-9　两因素完全随机嵌套实验的方差分析表

变异来源	SS	df	MS	F
A	60.67	2	30.335	19.482**
B(A)	14.67	6	2.445	1.57
误差	28.03	18	1.557	
合计	103.37	26		

方差分析结果表明，学习强度的效应是显著的，即增加练习强度对学习困难儿童学习生字有较大的帮助。结果还表明，不同训练时间的效应是不显著的。

（二）三因素完全随机嵌套实验设计

当实验中含有两个自变量和一个无关变量时，如何进行嵌套实验设计？我们举例来说明。研究者要探讨生字的结构特征和语义特征对于儿童学习和记忆的影响。实验中包括两个研究者感兴趣的自变量：每一个字表中的生字的结构特征（A 因素）和语义特征（B 因素），每个因素有两个水平。结构特征有繁（A1）、简（A2）两个水平；语义特征有熟悉（B1）、生僻（B2）两个水平。实验中还有一个因素是字表中汉字的系列顺序（C 因素）。由于字表中汉字的系列顺序可能会影响儿童的学习和记忆，因而是一个需要控制的无关变量。研究者设计了 20 个不同的随机顺序，并将系列顺序嵌套在四个处理条件水平的结合中。这样实验中有三个变量：两个研究者感兴趣的自变量和一个嵌套控制变量，其中自变量 A 和 B 各有两个水平，无关变量 C 有 C1 至 C20 共 20 个水平。有 200 名儿童参加了实验，每个处理条件水平结合下都有 50 名儿童参加，按照不同的系列顺序分组，分为 5 组，每组 10 人。

表 12-10　嵌套设计的汉字特征和被试分配

结构特征（A）和语义特征（B）	A1B1	A1B2	A2B1	A2B2
刺激的随机顺序	C1			
	C2			
	C3			
	C4			
	C5			
		C6		
		C7		
		C8		
		C9		
		C10		

续表

结构特征（A）和语义特征（B）	A1B1	A1B2	A2B1	A2B2
			C11	
			C12	
			C13	
			C14	
			C15	
				C16
				C17
				C18
				C19
				C20

实验实施后各组儿童学习和记忆汉字数量的结果见表 12-11。表12-12中是在各种处理水平的结合中儿童学习和记忆汉字的数量。

表 12-11　各组儿童对不同特征的汉字的学习记忆数量

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
$n=10$																				
A1B1	55	60	58	76	44															
A1B2						39	38	37	39	38										
A2B1											91	96	75	78	84					
A2B2																57	42	43	54	63

表 12-12　不同结构特征（A）和意义特征（B）汉字的学习和记忆

	A1	A2	Σ
B1	293	424	717
B2	191	259	450
Σ	484	683	1167

为了方便起见，我们只根据表 12-11 和表 12-12 中接受不同系列顺序字表的儿童组对不同特征的汉字的学习和记忆数量总和计算平方和：

$$SS_A = \frac{\sum_{j=1}^{p} A^2}{qr_{(jk)}n} - \frac{T^2}{pqr_{(jk)}n}$$

$$= \frac{484^2 + 683^2}{2 \times 5 \times 10} - \frac{1\ 167^2}{2 \times 2 \times 5 \times 10}$$

$$= 7\ 007.45 - 6\ 809.445$$

$$= 198.005$$

其中 A 是每个结构特征条件下汉字学习数量的总和，T 是所有特征条件下汉字学习数量的总和，n 是每组的被试数量，p 是结构特征的水平数量，q 是意义特征的水平数量，$r_{(jk)}$ 是每种处理水平结合中的汉字系列顺序的种类数。

$$SS_B = \frac{\sum_{k=1}^{q} B^2}{qr_{(jk)}n} - \frac{T^2}{pqr_{(jk)}n}$$

$$= \frac{717^2 + 450^2}{2 \times 5 \times 10} - \frac{1\ 167^2}{2 \times 2 \times 5 \times 10}$$

$$= 7\ 165.89 - 6\ 809.445$$

$$= 356.445$$

其中 B 是每个意义特征条件下汉字学习数量的总和。

$$SS_{AB} = \sum_{j=1}^{p} \frac{\sum_{k=1}^{q} AB^2}{r_{(k)}n} - \frac{T^2}{pqr_{(k)}n} - SS_A - SS_B$$

$$= \frac{293^2 + 424^2 + 191^2 + 259^2}{5 \times 10} - \frac{1\ 167^2}{2 \times 2 \times 5 \times 10} - 198.005 - 356.445$$

$$= 7\ 383.74 - 6\ 809.445 - 198.005 - 356.445$$

$$= 19.845$$

其中 AB 是每个结构特征和意义特征结合条件下汉字学习数量的总和。

$$SS_{C(AB)} = \sum_{j=1}^{p} \frac{\sum_{k=1}^{q} \sum_{l=1}^{r_{(jk)}} (ABC)^2}{n} - \sum_{j=1}^{p} \frac{\sum_{k=1}^{q} (AB)^2}{r_{(k)}n}$$

$$= \frac{55^2 + 60^2 + \cdots + 54^2 + 63^2}{10} - \frac{293^2 + 424^2 + 191^2 + 259^2}{5 \times 10}$$

$$= 7\ 500.9 - 7\ 383.74$$

$$= 117.16$$

其中 ABC 是每个字表的儿童组在某个结构特征和意义特征结合条件下汉字学习数量的总和。$SS_{C(AB)}$ 是总变异中由嵌套在 A、B 因素中的 C 因素引起的变异，在嵌套设计中是一种无关变异。

$$SS_{总变异} = \sum_{j=1}^{p}\sum_{k=1}^{q}\sum_{l=1}^{r_{(jk)}}\sum_{i=1}^{n}(ABCS)^2 - \frac{T^2}{pqr_{(jk)}n}$$
$$= 8\ 438.18 - 6\ 809.445$$
$$= 1\ 628.735$$

其中 ABCS 是每个被试在某个字表、汉字结构特征和意义特征结合条件下汉字学习数量的原始分数。已知被试的原始分数的总平方和是 8 438.18。

$$SS_E = SS_{总变异} - SS_A - SS_B - SS_{AB} - SS_{C(AB)}$$
$$= 1\ 628.735 - 198.005 - 356.445 - 19.845 - 117.16$$
$$= 937.28$$

表 12-13　三因素完全随机嵌套实验的方差分析表

变异来源	SS	df	MS	F
A	198.005	1	198.005	38.03**
B	356.445	1	356.445	68.45**
A×B	19.845	1	19.845	3.81*
C(AB)	117.16	16	7.323	1.41
误差	937.28	180	5.207	
合计	1 628.735	199		

方差分析的结果表明，汉字的结构特征和语义特征的主效应是显著的，两个因素的交互作用也是显著的，即生字的结构特征和语义特征对于儿童的学习和记忆有重要影响。实验中每个字表的汉字的系列顺序的效应是不显著的。

嵌套实验设计的重要用途在于，它使研究者可能在自然状态下利用固定团体进行某些研究，同时可以通过实验设计与统计分离出固定团体之间的差异带来的变异，从而使研究者的探索范围可扩展到一些传统实验设计无法涉及的领域，同时，又不降低研究的精度。嵌套实验设计还可以用于分离实验中的无关变量带来的变异。使用嵌套设计的前提是假设自变量与无关变量之间没有交互作用。如果事先我们知道它们之间存在交互作用，

这种实验设计是不合适的。读者如果希望更深入了解无关因素在实验处理条件中的嵌套的数据分析方法的原理，还可以将本节的内容与《心理与教育研究中的多因素实验设计》（舒华，1994）一书中有关固定被试团体在实验处理条件中的两因素嵌套设计和三因素嵌套设计的数据分析方法相比较。近年来还发展了一些新的统计方法，如多层分析（multilevel analysis），来解决多层嵌套数据的问题（张雷等，2002）。

第二节 协方差分析

前面的章节中介绍了许多如何通过实验设计来减少实验误差，以便获得对处理效应的无偏估计的方法。实验控制能够以各种方式减少实验误差，如随机分配被试给各实验处理条件，对被试进行匹配以形成同质区组，改善因变量测量等。减少实验误差的另一个途径是使用统计控制，通过统计分离出实验中潜在的偏差源，这些偏差是很难或者不可能通过实验控制消除的。

本节介绍的一种统计控制是协方差分析（analysis of covariance），它将回归分析和方差分析相结合，可以测量除自变量外的一个或多个协变量的影响。协变量指在实验中未被控制，但被认为是影响因变量的变异源。通过协方差分析，可以分离出协变量的影响。这种方法的优点是：可以通过分离协变量的影响减少实验误差，增加检验力；同时，减少由无关变量引起的实验单元间差异导致的偏差。

一、协方差的应用

协方差分析主要是提供一种对实验结果的调整。在这种实验中，被试之间事先存在的差异或实验处理组之间在其他方面事先存在的差异可能会影响实验的结果，因此调整这些差异，对减少误差变异，提高实验的敏感性是非常必要的。协方差分析对实验结果调整程度的大小依赖于选择的协变量与因变量之间的相关。

协方差分析经常用于以下的研究情景。

（1）有些研究中使用固定的被试组，如在教育、工业研究中经常有这

样的情况：将固定的被试团体分配给不同的实验处理条件。例如，要研究四种教学方法的效应，为了让实验结果更接近自然，一般不是将学生随机分配给不同的教学条件，而是在四个自然的班级中进行不同教学方法的实验。如果在教学之前，四个班级之间存在学习能力或其他类似特征的差异，这些额外的变量会影响对处理效应的估计。

（2）在实验中，虽然被试是随机分配的，但各实验处理组的被试在实验开始时仍然可能在某个无关变量上存在差异。例如，一个实验要估计不同药物对白鼠的刺激泛化的效应。在实验开始时，白鼠被随机分配进各个实验组，让它们学习按键反应。如果各组需要不同的训练次数以建立稳定的按键反应，表明各组之间存在着学习能力的不同。实验者可能在研究结论中发现，刺激泛化的量与建立稳定的按键反应需要的训练量是相关的，各组白鼠的刺激泛化分数可能是受学习能力影响的。

（3）除自变量、被试变量之外，可能还存在另外的变量影响因变量结果，这些变量是在实验开始以后出现的。例如，在四种教学方法的研究中，还有些变量可能影响对因变量的测量，如四个班级的学生在本班教室学习的小时数。学校的教室日常安排的差异可能对一个教室的学生比对另外一个教室的学生提供了更多的教室学习时间，这个变量也可能影响学习成绩。由于对学生的学习时间很难进行实验控制，一个可行的办法是记录下学生每天在教室的学习时间，然后利用这个信息在因变量中作适当调整。在这个例子中，教室学习时间这个伴随变量是在教学实验开始后出现，但在教学效果测验前起作用的，因而这个伴随变量也可能影响对实验结果的估计。

统计控制和实验控制在减少偏差、增加精度上不是互相排斥的，有时用实验控制很方便，而在另一些情况下用统计控制更加方便。一般来说，在可能的情况下首先考虑使用实验控制，统计控制通常是在很难或无法进行实验控制的情况下使用的。

二、协变量的选择

协方差分析中的协变量（concomitant variables）是需要小心选择的，重要的原则是：协变量调整消除的变异应当与自变量的变化无关。协方差分

析可以与我们介绍过的实验设计相结合。选择协变量要考虑的因素如下。

（1）实验中含一个或多个被认为影响因变量但与自变量的变化无关的额外变异源。

（2）对这些额外变异源进行实验控制是不可能的。

（3）有可能测量这个（些）无关变异，它（们）独立于处理效应。独立于处理效应是指以下情况：①无关变量是在处理水平实施之前获得的；②无关变量的观察是在实验处理实施后，但在实验处理影响因变量前获得的；③可以假定无关变量是不受实验处理影响的。例如，在上述的四种教学方法的研究中，选择将教室的学习时间设为协变量的前提是：假设各班教室学习时间量与分配的教学方法是相对独立的。

协变量选择的主要标准是：协变量与因变量高相关。在多数情况下，协变量的分数是在实施实验处理前获得的。例如，可以对所有实验参加者实施一个某种类型的前测以获得协变量的分数，或者协变量的分数可能来自被试已经有的一些记录，如成就分数、智力测验分数、年级成绩平均等是常用的协变量分数。一些相对恒定的被试能力，如阅读能力、短时记忆能力，常常被选做协变量。因为我们假设这些相对恒定的被试能力（如阅读能力）是不受实验处理（如教学方法）的影响的。但将被试的一些暂时倾向，如焦虑、自我等，作为协变量，可能是不合适的。协方差分析结论的正确性依赖于协变量与实验处理的信息的独立性，因此下结论要非常小心。

还有一些情况下，协变量的分数是在实验处理实施后，但在实验处理影响因变量前获得的，这只有当确认实验处理和协变量分数相对独立时才是可行的。如上面所举的研究四种教学方法的例子，选取四个班级的学生在教室学习的小时数作为协变量，记录学生每天的学习时间是在教学开始后，但在因变量测量之前进行的。能够利用这个信息在因变量中作调整，是基于假设学生在教室的学习时间量是不受学生所分配的实验处理——教学方法的类型——的影响的。

在实验过程中采集的协变量分数的合适性是一个非常重要的问题，重要的是要保证协变量的变异是独立于处理效应的。我们再举两个记忆研究实验的例子，其中一个研究中选取的协变量是合适的，而另一个研究中选取的协变量是不合适的。在第一个研究中，研究者探讨在若干恒定的训练

后的遗忘过程。选择五个独立的被试组，各进行 10 次训练。然后在训练结束 2 分钟、20 分钟、1 小时、6 小时、24 小时后进行测验。研究者发现使用恒定的训练次数，在 10 次训练后不是所有的被试达到同样的行为水平。前人的研究表明，训练结束时的学习程度与后来的延迟回忆量是正相关的。由于被试在固定次数训练后会存在学习程度上明显的差异，最终的学习分数或学习程度可以作为遗忘过程研究中的一个协变量。我们假设所有的被试在实验实施前水平是相同的。

在另一个研究中，假定五个独立被试组接受难度不同材料的训练，各进行 10 次训练。训练结束后，所有被试在 24 小时后接受测验。与第一个研究不同的是，五组被试学习的材料是难度不同的。这时，有两个变量会影响 10 次训练后被试达到的行为水平：一个是材料的难度，另一个是被试的个体差异。因此，不同被试在学习程度上的差异既包含了组内的差异，又包含组间的差异。虽然已知被试学习后的行为差异与延迟回忆的分数是高相关的，但在这种情况下，仍然将最终的学习分数或学习程度作为协变量进行协方差分析是不合适的。因为作为协变量的最终学习分数或学习程度，不仅与被试的个体差异有关，而且与实验处理有关，通过协方差分析调整这种系统差异是不可能的。

在教育研究中，分配学校中不同班级的学生接受不同的实验处理是更方便的，但是这些固定的被试团体经常在一些与实验中因变量有关的重要特征上是有差异的。如各班级的智力分数、年级平均分数等经常是不相同的。通常情况下，这些被试的有关信息在实验开始前就是存在的，因此协方差分析是一个可能校正各被试组差异的有效方法。然而在很多情况下，协方差分析是不适用的。例如，当这些固定团体原来不是随机分配的，或他们接受的实验处理不相同时，协方差分析就是不适用的。

三、协方差分析的原理

协方差分析是通过相应的协变量分数，对因变量分数进行的统计调整。有两种基本的调整：对每一个实验处理组中被试的无关变量的偶然误差的调整；对实验处理组的偶然误差的调整。第一个调整是针对个体的，基于每个被试偏离其所在被试组平均数的标准差。第二个调整对于一个被

试组内的所有被试是恒定的，基于每个被试组平均数偏离总平均数的标准差（Kirk，1982）。

在下列情况下，协方差分析是十分有用的。

（1）调整误差项：假定协变量的组平均是相等的，但组内被试之间有差异，在这种情况下，协方差分析会减小误差项的大小，使处理组之间的差异更明显，产生一个更大的 F 值。当然，随机分配给各实验处理条件的被试组获得相同平均的可能性是零，这样除了对个体误差在协变量上误差项的校正，协变量还校正了处理组的差异。

（2）当协变量和因变量的平均差异方向相同时的调整：在有些情况下，某些实验处理的差异包含了实验开始之前已存在的差异，对这些差异的调整将导致减小处理效应的平方和。

（3）当协变量和因变量的平均差异方向相反时的调整：在另一些情况下，由于处理效应预期被试组在因变量的一个方向上有差异，而协变量预期被试组在因变量的另一个方向上有差异，从而导致因变量测量的处理效应被减小。在这种情况下，调整会加大处理效应的平方和，即当分离了实验前存在的差异，实验处理表现出是更有效的。

协方差分析的调整的基本逻辑是：协方差模型的关键是线性回归系数，系数指出了协变量和因变量相关的直线斜率，回归线反映了可以从协变量相应分数预测的因变量分数。在方差分析中，离散分数形成计算的基本成分。在协方差分析中，我们仍然对离散分数感兴趣，但这时是残差标准差，它是与控制变量没有相关的变异，即与控制变量的线性效应无关的变异。协方差分析可以看成方差分析的一个特例，其中对因变量的测量分数作了调整，以解决在某些已知与因变量相关的预期变量上存在个体差异的问题。

协方差分析的实验设计模型：

$$Y_{ij}=\mu+\alpha_j+\beta_w(X_{ij}-\overline{X_{..}})+\varepsilon_{i(j)}$$

$$(i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p)$$

其中 Y_{ij} 是被试 i 在处理水平 j 中的分数，μ 是总体平均数，α_j 是水平 j 的处理效应，β_w 是线性回归系数，X_{ij} 是被试 i 在处理水平 j 中的协变量分数，$\overline{X_{..}}$ 是协变量的总平均数，$\varepsilon_{i(j)}$ 是误差变异。

我们举例来说明协方差分析的使用。研究者想探讨儿童在学习多个汉

字家族时，是分散学习好还是集中学习好。所谓的汉字家族，是指含有同一个声旁的形声字所组成的集合，如果这个声旁能够独立成字，则该声旁也是这个家族的成员之一，如“伴”“绊”“拌”“判”“半”同属一个家族。分散学习指的是把属于多个家族中的多个汉字随机呈现，让学生学习。集中学习是指把属于同一个家族的汉字放在一起让学生进行学习，学完了一个家族的汉字，再学习另一个家族的汉字。由于客观教学实践的限制，这两种教学方法只能在两个自然班级中进行，每个班随机接受一种教学方法。在教学结束后，进行一个完全相同的汉字注音测验，以考察教学效果。虽然所选择的汉字及其家族全部都是儿童没有学习过的，但考虑到儿童原有的识字量可能会影响测验结果，在实验开始前，对每个学生进行了一个识字量测验。

这是单因素的被试间实验设计，教学方法（A）是自变量，有两个水平（分散学习和集中学习），识字量测验的分数（X）是协变量，汉字注音测验的分数（Y）是因变量。为了计算方便，假设每个自然班中有八个学生。我们将分别通过手工计算和 SPSS 软件计算说明协方差的分析方法。

四、协方差分析的手工计算

为了更清楚地介绍协方差分析的计算过程（Kirk，1982），我们根据表 12-14 中每个被试的原始分数进行计算。

表 12-14　原始数据表

	A_{X1}	A_{Y1}	A_{X2}	A_{Y2}
	10	15	7	14
	6	1	8	9
	5	4	7	16
	8	6	3	7
	9	10	6	13
	4	0	8	18
	9	7	6	13
	12	13	8	6
Σ	63	56	53	96

其中 A_{X1} 和 A_{X2} 是每个学生的识字量测验分数，A_{Y1} 和 A_{Y2} 是每个学生的汉字注音测验的分数。

首先根据原始分数计算各种总和：

$$\sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij} = 15+1+4+\cdots+6 = 152$$

$$\frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij}\right)^2}{np} = \frac{152^2}{8\times 2} = 1\ 444$$

$$\sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij}^2 = 15^2+1^2+\cdots+6^2 = 1\ 876$$

$$\sum_{j=1}^{p}\frac{\left(\sum_{i=1}^{n}Y_{ij}\right)^2}{n} = \frac{56^2}{8}+\frac{96^2}{8} = 1\ 544$$

$$\sum_{i=1}^{n}\sum_{j=1}^{p}X_{ij} = 10+6+\cdots+8 = 116$$

$$\frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{p}X_{ij}\right)^2}{np} = \frac{116^2}{8\times 2} = 841$$

$$\sum_{i=1}^{n}\sum_{j=1}^{p}X_{ij}^2 = 10^2+6^2+\cdots+8^2 = 918$$

$$\sum_{j=1}^{p}\frac{\left(\sum_{i=1}^{n}X_{ij}\right)^2}{n} = \frac{63^2}{8}+\frac{53^2}{8} = 847.25$$

$$\frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{p}X_{ij}\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij}\right)}{np} = \frac{116\times 152}{8\times 2} = 1\ 102$$

$$\sum_{i=1}^{n}\sum_{j=1}^{p}X_{ij}Y_{ij} = 10\times 15+6\times 1+\cdots+6\times 8 = 1\ 184$$

$$\sum_{j=1}^{p}\frac{\left(\sum_{i=1}^{n}X_{ij}\right)\left(\sum_{i=1}^{n}Y_{ij}\right)}{n} = \frac{63\times 56}{8}+\frac{53\times 96}{8} = 1\ 077$$

然后进行各种平方和的计算：

$$T_{YY}=\sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij}^2-\frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij}\right)^2}{np}=1\ 876-1\ 444=432$$

其中 T_{YY}是未调整的因变量的总平方和。

$$A_{YY}=\sum_{j=1}^{p}\frac{\left(\sum_{i=1}^{n}Y_{ij}\right)^2}{n}-\frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij}\right)^2}{np}=1\ 544-1\ 444=100$$

其中 A_{YY}是未调整的组间平方和。

$$S_{YY}=\sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij}^2-\sum_{j=1}^{p}\frac{\left(\sum_{i=1}^{n}Y_{ij}\right)^2}{n}=1\ 876-1\ 544=332$$

其中 S_{YY}是未调整的组内平方和。

$$T_{XX}=\sum_{i=1}^{n}\sum_{j=1}^{p}X_{ij}^2-\frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{p}X_{ij}\right)^2}{np}=918-841=77$$

其中 T_{XX}是协变量的总平方和。

$$A_{XX}=\sum_{j=1}^{p}\frac{\left(\sum_{i=1}^{n}X_{ij}\right)^2}{n}-\frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{p}X_{ij}\right)^2}{np}=847.25-841=6.25$$

其中 A_{XX}是协变量的组间平方和。

$$S_{XX}=\sum_{i=1}^{n}\sum_{j=1}^{p}X_{ij}^2-\sum_{j=1}^{p}\frac{\left(\sum_{i=1}^{n}X_{ij}\right)^2}{n}=918-847.25=70.75$$

其中 S_{XX}是协变量的组内平方和。

$$T_{XY}=\sum_{i=1}^{n}\sum_{j=1}^{p}X_{ij}Y_{ij}-\frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{p}X_{ij}\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij}\right)}{np}=1\ 184-1\ 102=82$$

其中 T_{XY}是协变量和因变量的共变平方和。

$$A_{XY}=\sum_{j=1}^{p}\frac{\left(\sum_{i=1}^{n}X_{ij}\right)\left(\sum_{i=1}^{n}Y_{ij}\right)}{n}-\frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{p}X_{ij}\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{p}Y_{ij}\right)}{np}$$

$$=1\ 077-1\ 102=-25$$

其中 A_{XY}是组间共变平方和。

$$S_{XY}=\sum_{i=1}^{n}\sum_{j=1}^{p}X_{ij}Y_{ij}-\sum_{j=1}^{p}\frac{\left(\sum_{i=1}^{n}X_{ij}\right)\left(\sum_{i=1}^{n}Y_{ij}\right)}{n}=1\ 184-1\ 077=107$$

其中 S_{XY}是组内共变平方和。

$$T_{adj}=T_{YY}-\frac{T_{XY}^2}{T_{XX}}=432-\frac{82^2}{77}=344.68$$

其中 T_{adj}是调整后的因变量的总平方和。

$$A_{adj}=T_{adj}-S_{adj}=344.68-170.18=174.5$$

其中 A_{adj}是调整后的因变量的组间平方和。

$$S_{adj}=S_{YY}-\frac{S_{XY}^2}{S_{XX}}=332-\frac{107^2}{70.75}=170.18$$

其中 S_{adj}是调整后的因变量的组内平方和。

下面我们通过对同一组数据分别进行协方差分析和普通的方差分析，来对比一下两种统计方法的异同。首先，如果我们使用识字量测验的成绩为协变量来调整因变量分数，协方差分析发现两种教学方法之间的差异是非常显著的（见表 12-15）。表中 A_{adj}是用协变量调整后的组间变异，S_{adj}是用协变量调整后的组内变异。

表 12-15　协方差分析表

变异来源	SS	df	MS_{adj}	F
组间（A_{adj}）	174.5	$p-1=1$	174.5	13.33**
组内（S_{adj}）	170.18	$np-p-1=13$	13.091	
合计	344.68	$np-2=14$		

然而，如果使用未调整的因变量分数，简单的方差分析发现两种教学方法之间的差异仅仅达到边缘显著（见表 12-16）。表中 A_{YY}是未调整的组间变异，S_{YY}是未调整的组内变异。

表 12-16　方差分析表

变异来源	SS	df	MS	F
组间（A_{YY}）	100	$p-1=1$	100	4.217
组内（S_{YY}）	332	$p(n-1)=14$	23.714	
合计	432	$np-1=15$		

五、协方差分析的 SPSS 软件计算

（一）数据格式

在上述探讨儿童汉字学习方法影响的研究中，只包含一个自变量，即教学方法。它是一个被试间变量（用 A 表示），有两个水平（用 1 和 2 表示，分别代表分散学习和集中学习），识字量测验的分数是协变量（用 X 表示），汉字注音测验的分数是因变量（用 Y 表示）。在 SPSS 数据窗口中的格式如下：

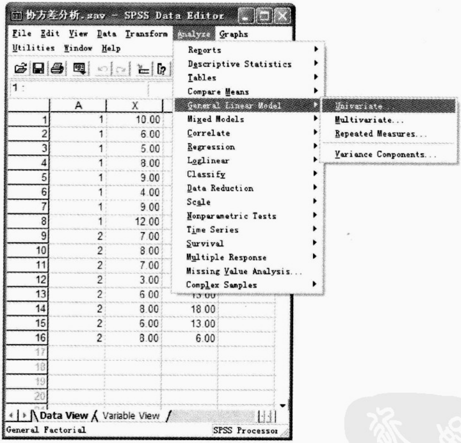
	A	X	Y
1	1	10.00	15.00
2	1	6.00	1.00
3	1	5.00	4.00
4	1	8.00	6.00
5	1	9.00	10.00
6	1	4.00	.00
7	1	9.00	7.00
8	1	12.00	13.00
9	2	7.00	14.00
10	2	8.00	9.00
11	2	7.00	16.00
12	2	3.00	7.00
13	2	6.00	13.00
14	2	8.00	18.00
15	2	6.00	13.00
16	2	8.00	6.00

需要注意的是，由于所有的被试都参加了识字量测验和汉字注音测验，因此，每个被试都有两个成绩，应该安排在同一行。而教学方法是被

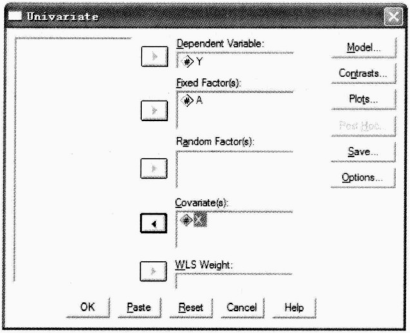
试间变量，各有 8 个学生参加了分散学习和集中学习，因此，共有 16 个被试参加实验，一共有 16 行数据。

（二）数据分析

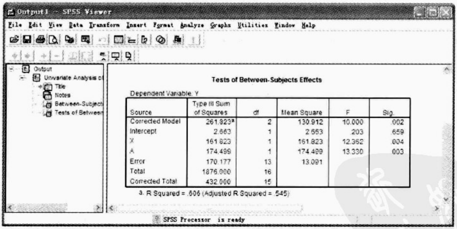
激活 Analyze 菜单，选 General Linear Model 中的 Univariate... 命令项。



弹出 Univariate 对话框，在对话框左侧的变量列表中，选中变量 A，点击 ▸ 钮，使之进入 Fixed Factor(s) 框；选中变量 X，点击 ▸ 钮，使之进入 Covariate(s) 框；选中变量 Y，点击 ▸ 钮，使之进入 Dependent Variable 框。



点击 OK 钮，开始协方差分析。输出结果如下。



Tests of Between-Subjects Effects

Dependent Variable: Y

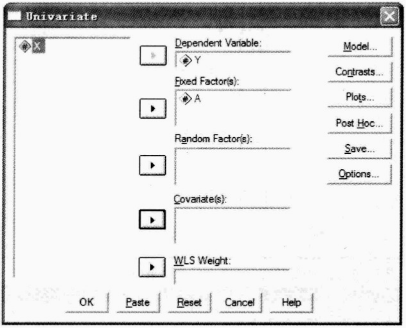
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	261.823ª	2	130.912	10.000	.002
Intercept	2.663	1	2.663	.203	.659
X	161.823	1	161.823	12.362	.004
A	174.499	1	174.499	13.330	.003
Error	170.177	13	13.091		
Total	1876.000	16			
Corrected Total	432.000	15			

a. R Squared = .606 (Adjusted R Squared = .545)

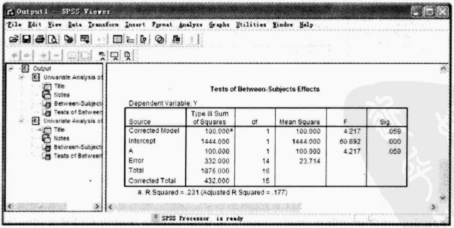
图 12-1 使用调整的因变量分数的协方差分析输出结果

这一结果与我们手工计算的结果完全一致，显示在使用识字量测验的成绩为协变量来调整因变量分数时，两种教学方法之间的差异是非常显著的，$F(1，13)=13.330$，$p=0.003$。

下面我们对同样的数据进行不使用识字量测验的成绩作为协变量的方差分析。在 Univariate 对话框，只选中变量 A，点击 ▶ 钮进入 Fixed Factor(s) 框；选中变量 Y，点击 ▶ 钮进入 Dependent Variable 框。



点击 OK 钮，则得到如下输出结果。



Tests of Between-Subjects Effects

Dependent Variable: Y

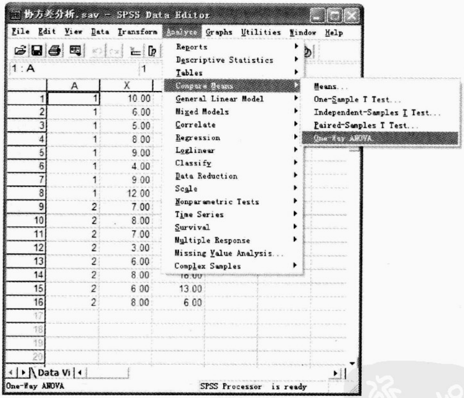
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	100.000ᵃ	1	100.000	4.217	.059
Intercept	1444.000	1	1444.000	60.892	.000
A	100.000	1	100.000	4.217	.059
Error	332.000	14	23.714		
Total	1876.000	16			
Corrected Total	432.000	15			

a. R Squared = .231 (Adjusted R Squared = .177)

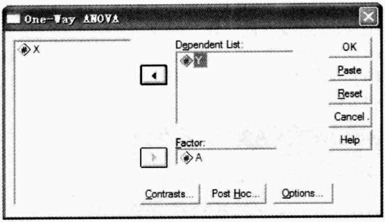
图 12-2　使用未调整的因变量分数的协方差分析输出结果

这一结果与我们手工计算的结果也完全一致，显示如果不把识字量测验的成绩作为协变量，而使用未调整的因变量分数，则两种教学方法之间的差异仅仅达到边缘显著 $F(1,\ 14)=4.217$，$p=0.059$。

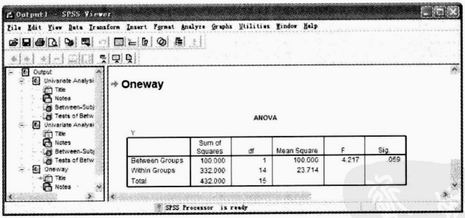
在没有考虑识字量测验作为协变量时，这个实验仅是一个单因素两水平的被试间设计，因此，可以直接使用 Compare Means 中的 One-Way ANOVA 分析。



原始数据输入时，将汉字注音测验的成绩 Y 作为因变量放入 Dependent List 中，将教学方式 A 作为自变量放入 Factor 中。



点击 OK，得到的结果与前面不使用协变量 X 得到的结果完全一致，两种教学方法之间的差异仅仅达到边缘显著 $F(1, 14)=4.217$，$p=0.059$。



	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	100.000	1	100.000	4.217	.059
Within Groups	332.000	14	23.714		
Total	432.000	15			

图 12-3　方差分析输出结果

以上举例说明了 One-Way ANOVA 分析和 General Linear Model 分析的基本数学原理是一样的，只不过 One-Way ANOVA 分析是 General Linear Model 分析的一个特例，适用于只有一个自变量的情况。

（三）协变量和因变量的平均差异的不同方向的调整举例

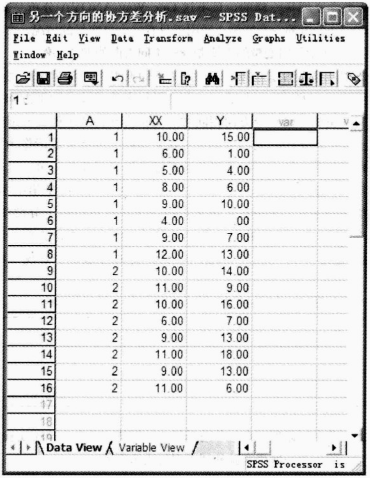
在协方差分析的原理部分提到协变量的调整有两种情况：当协变量和因变量的平均差异方向相反时，调整会加大处理效应的平方和；当协变量和因变量的平均差异方向相同时，调整将导致减小处理效应的平方和。在上面的例子中，我们可以看到在使用识字量测验成绩作为协变量之后，两种教学方法之间的差异从边缘显著变成了非常显著。这是由于在进行实验之前，分散学习班学生的识字量测验成绩高于集中学习班，而在教学干预之后，集中学习班学生的汉字注音成绩高于分散学习班，协变量 X 和因变量 Y 的平均差异方向是相反的。两个班在两个测验上的平均数见表 12-17，其中 X 是识字量测验的分数，Y 是汉字注音的分数。

表 12-17　两班学生的识字量测验和汉字注音的分数

	X	Y
A1	7.875	7.0
A2	6.875	12.0

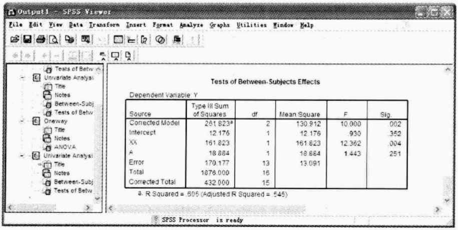
因此，我们可以理解在这种情况下，使用协变量分离出了实验之前存在的反方向的差异，使得实验处理表现得更为有效。

在另一种情况下，协变量和因变量的平均差异方向是相同的。例如，我们仍使用上例中的数据，只是把每个集中学习班学生的识字量测验成绩加上 3（用 XX 表示），使得集中学习班学生的识字量测验成绩和汉字注音成绩都高于分散学习班，数据如下：



	A	XX	Y
1	1	10.00	15.00
2	1	6.00	1.00
3	1	5.00	4.00
4	1	8.00	6.00
5	1	9.00	10.00
6	1	4.00	.00
7	1	9.00	7.00
8	1	12.00	13.00
9	2	10.00	14.00
10	2	11.00	9.00
11	2	10.00	16.00
12	2	6.00	7.00
13	2	9.00	13.00
14	2	11.00	18.00
15	2	9.00	13.00
16	2	11.00	6.00

使用与上例完全相同的协方差分析，即教学方法是一个被试间变量（用 A 表示），有两个水平（用 1 和 2 表示，分别代表分散学习和集中学习），识字量测验的分数是协变量（用 XX 表示），汉字注音测验的分数是因变量（用 Y 表示），得到以下结果。



Tests of Between-Subjects Effects

Dependent Variable: Y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	261.823[a]	2	130.912	10.000	.002
Intercept	12.176	1	12.176	.930	.352
XX	161.823	1	161.823	12.362	.004
A	18.884	1	18.884	1.443	.251
Error	170.177	13	13.091		
Total	1876.000	16			
Corrected Total	432.000	15			

a. R Squared = .606 (Adjusted R Squared = .545)

图 12-4 协方差分析的输出结果

从图 12-4 中我们看到，$F(1, 13)=1.443$，$p=0.251$，两种教学方式之间的差异连边缘水平都达不到了。其原因在于，当把每个集中学习班学生的识字量测验成绩加上 3 后，集中学习班学生在教学前的识字量测验成绩和教学后的汉字注音成绩都高于分散学习班。在这种情况下，实验处理的差异包含了实验开始之前已存在的差异，使用协变量分离出了实验之前就已经存在的同方向的差异，使得实验处理的效果变得不明显了。

（四）协方差分析的调整原理举例

我们再举一个例子说明协方差分析的调整原理。研究者试图比较两种对儿童进行早期干预方法的有效性。首先，随机分配两组儿童进入两个干预项目，每组 10 个儿童。在进行干预前和实施干预后，对两组儿童分别进行了智力测验。在这个实验中，因变量是实施干预后的智力测验分数（后测分数）。考虑到儿童可能在原有智力测验上存在差异，将干预前的智力测验分数（前测分数）作为协变量。实验得到的原始数据如下：

表 12-18 两组儿童智力测验的原始数据

被试	干预方法	前测分数	后测分数
1	1	80	87
2	1	125	124
3	1	103	105
4	1	101	107
5	1	125	93
6	1	89	93
7	1	111	93
8	1	116	127
9	1	110	118
10	1	95	106
11	2	101	103
12	2	125	121
13	2	105	109
14	2	104	116
15	2	111	123
16	2	110	139
17	2	125	135
18	2	68	101
19	2	101	121
20	2	95	97

我们用两种方法对数据进行计算。第一种方法，不考虑儿童在原有智力测验上可能存在的差异，只用实施干预后的智力测验分数（后测分数）作为因变量，进行方差分析检验（见表 12-19）。结果表明，两种干预方法的效应是不显著的，$F(1, 18)=3.17$，$p>0.05$。

表 12-19　方差分析表（基于后测分数）

来源	SS	df	MS	F	p
干预	627.20	1	627.20	3.17	0.092
误差	3 564.60	18	198.03		

第二种方法，我们考虑儿童可能在原有智力测验上存在的差异，将干预前的智力测验分数（前测分数）作为协变量，进行协方差分析检验（见表 12-20）。结果表明，两种干预方法的效应是显著的，$F(1, 17)=5.28$，$p<0.05$。

表 12-20　协方差分析表（用前测分数做协变量）

来源	SS	df	MS	F	p
干预	690.23	1	690.23	5.28	0.035
回归	1 342.41	1	1 342.41		
误差	2 222.19	17	130.72		

为什么两种计算得出的结论会不相同？让我们比较一下两个实验设计模型。

方差分析的实验设计模型是：

$$Y_{ij}=\mu+\alpha_j+\varepsilon_{i(j)}$$

$$(i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p)$$

协方差分析的实验设计模型是：

$$Y_{ij}=\mu+\alpha_j+\beta_w(X_{ij}-\overline{X}_{..})+\varepsilon_{i(j)}$$

$$(i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p)$$

可以看出，与基于后测分数的方差分析相比，在协方差分析中，前测分数与后测分数关系的调整导致一部分可以从协变量预测因变量的变异被从误差变异中分离出来，减小了误差变异，使干预的效应显示出来。

本章主要观点

• 嵌套实验设计经常适用的两种情况：一种是被试组在实验处理条件中的嵌套，实验设计的目的主要是分离出团体效应；另一种是控制因素或无关因素在实验处理条件中的嵌套，实验设计旨在分离出与实验处理效应

相混淆的无关变量的效应。

• 嵌套实验设计的一个重要用途是使研究者可能在自然状态下利用固定团体进行某些研究，同时可以通过实验设计与统计分离出固定团体之间的差异带来的变异，从而使研究者的探索范围可扩展到一些传统实验设计无法涉及的领域，同时又不降低研究的精度。

• 嵌套实验设计也可以用于分离实验中的无关变量带来的变异。当实验中有些无关变量可能影响因变量的测量，并与实验处理效应相混淆时，通过无关变量在实验处理条件中的嵌套设计可以帮助分离出这样的无关变量的效应。

• 协方差分析是通过统计控制减少实验误差的一个途径，其基本方法是通过统计分离出实验中潜在的偏差源，而这些偏差是很难或者不可能通过实验控制消除的。

• 协变量指在实验中未被控制但被认为是影响因变量的变异源。协方差分析是将回归分析和方差分析相结合，测量除自变量外的一个或多个协变量的影响，并分离出协变量的效应。

• 选择协变量要考虑的因素包括：实验中含有一个或多个影响因变量但与自变量的变化无关的额外变异源，对这些额外变异源进行实验控制是不可能的，这个（些）无关变异是独立于处理效应的。协变量选择的主要标准是协变量与因变量高相关。

思考题

1. 什么是嵌套实验设计？嵌套设计与交叉设计的区别是什么？在什么情况下需要使用嵌套设计？

2. 将被试组嵌套在处理条件中的实验设计的主要目的是什么？其实验设计和方差分析方法是什么？

3. 将无关因素嵌套在处理条件中的实验设计的主要目的是什么？其实验设计和方差分析方法是什么？

4. 什么是协方差分析？在什么情况下需要使用协方差分析？

5. 什么是协变量？如何选择协变量？

6. 协方差分析的原理是什么？协方差分析和方差分析的差别是什么？

第十三章
实验数据的整理与处理

心理学是一门实验的科学，其重要特点是需要在收集数据的基础上获得结论。因此，客观、可靠的数据是心理学研究的根本。没有可靠的数据，所有的研究结论就失去了根基。本章中，我们将介绍一些常用的数据整理和处理的方法。

第一节　原始数据的整理

整理数据是了解研究结果、进行进一步统计的重要步骤。获得原始数据（raw data）后，研究者需要按照规范，或者说按照相应研究群体可接受的方式对数据进行整理和处理。这些规范可能包括：如何处理极端数据，如何对待“不符合假说”的数据等。收集的数据中经常可能会含有极端数据，即某些数据的数值远远大于或小于平均值。如何对待这样的极端数据？有些人会根据自己的经验去除一些“远远大于平均数”“远远小于平均数”或“不可能”的数值。然而，如果任意去除极端数据，不同研究者的标准可能是不同的，从而使实验之间的结果无法进行比较。又如，实验中还经常会发现一些“特别”的被试，他们的反应与其他被试的反应有很大的差别。例如，有些被试在一些反应时任务中犯错误的比率比其他被试高得多。哪些被试是“特别被试”，需要从总数据中去除他们的数据？如果研究者根据个人的经验任意去掉一些被试的数据，也可能因为研究者去除标准的不同，从而导致得到不同的结果和结论。以上两种“任意”处理数据的方式，很可能会导致实验结论的不同。下面我们介绍一些整理原始数据的方法。

一、极端数据的去除或替代

原则上，我们尽量使用所有的有效数据来进行统计分析。但是根据样本分布，在进行统计检验分析之前，往往需要先将那些被认为不是来自于该样本总体的数据剔除掉。在经验操作上，可以按照平均数加减三个标准差的原则来处理。在这个范围之外的数据，我们称之为极端数据，统计上有理由认为它不是来自抽样样本所代表的总体，需要对这部分数据进行特殊的处理。

处理的方法有两个：一个是用平均数替代；另一个是用上限或下限对应的边缘数值替代。这两种替代的假设是不同的。前者假设该极端数据不反映任务难度、加工深度等任何信息，仅仅是一个“意外”，这种情况下就可以用样本平均数来代替。而后者则认为该极端数据传达了一定的信息，可能反映了加工难度太高或者太低等。获得的所有数据是一个连续体，只是该极端数据远远偏离了平均数，但是偏离的方向和距离是有意义的。这时可以用上限或下限对应的边缘数值来替代。

在 Excel 中，可以利用提供的 if 函数及其强大的复制、粘贴功能来实现这两类替代。例如，在一个反应时实验中，这两种替代都需要先计算某条件、某项目或某被试的相应的总的反应时平均数和标准差（不包括错误反应数据），然后计算标准所对应的上下限。需要注意的是：在 Excel 中的数据计算，空的单元格不参与，也可以理解为是用平均数替代。所以，如果用平均数替代的话，只需要把极端值替换为空的单元格即可。需要判断一个数值是否在范围上下限之内时，可用函数“＝if（or（b3＞上限，b3＜下限），“”，b3)”。如果用边缘数据替代，就用相应标准（比如三个标准差）对应的上下限数据来替代，即用函数“＝if(b3＞上限，上限，b3)”和“＝if（b3＜下限，下限，b3)”。

二、观察描述统计的结果

在进行统计检验之前，对研究者非常重要的是观察一组数据的分布情况以及平均数、标准差、基本相关等描述统计数据，这可以使研究者对应当如何进一步处理数据做到心中有数。有经验的研究者能从平均数和标准差上初步了解各组数据的差异模式，对统计检验的结果有一个初步的

估价。

描述统计的结果是基于有效实验数据的。数据有效是相对于无效而言的。就常用的反应时数据而言，数据无效包括两个方面：一类是错误反应，因为错误反应和正确反应的心理加工过程可能是不同的，因此一般的反应时研究中所计算的有效数据只包括正确反应的反应时；另一类则是上面提到的极端数据，虽然是正确反应，但是它代表的心理含义可能与所要探讨的问题不同。所以，反应时研究中的描述统计是在进行了错误反应和极端数据处理后得到的。

在 Excel 中计算一组数据的平均数可以使用公式"=average(*)"，标准差计算可以用公式"=stdev(*)"，求中数用公式"=median(*)"，求众数用公式"=mode(*)"等。其中，* 表示数据区。当然，也可以用 SPSS 中描述统计功能得到这些信息。

观察描述统计结果有时候会告诉我们许多有用的信息。例如，我们做实验时经常会先做一个预实验，实际上是先测试一个小样本。如果预实验的结果表明实验设计是可行的话，再施测更多的被试，继续完成实验。如何通过观察小样本数据的描述统计特征估价实验设计是否可行？我们来看看下面的例子。

在一个启动实验（实验 1）中，有三种启动条件和一个控制条件，研究者先测试了 30 名被试。被试反应时的描述统计结果如下：

表 13-1　30 名被试在启动实验 1 中反应时数据（单位：ms）

	被试数	平均数	标准差	启动效应
启动 1	30	632	84	+3
启动 2	30	633	83	+2
启动 3	30	633	82	+2
控制	30	635	69	

从表 13-1 中的启动实验 1 的反应时平均数结果可以看出，各种启动条件下的启动效应应当是不显著的。而且由于各种条件下的标准差差异较小，说明数据是比较稳定的，通过增加被试获得显著的启动效应的

可能性也比较小。在这样的情况下，测试更多的被试可能是没有效果的。

我们再看另外一个例子。在一个启动实验（实验 2）中，有三种启动条件和一个控制条件，研究者先测试了 34 名被试，被试反应时的描述统计结果如下。

表 13-2　34 名被试在启动实验 2 中反应时数据（单位：ms）

	被试数	平均数	标准差	启动效应
启动 1	34	761	130	+3
启动 2	34	760	136	+4
启动 3	34	775	125	−11
控制	34	764	120	

从表 13-2 中启动实验 2 的反应时平均数结果可以看出，启动 3 条件下有一定的抑制效应，虽然抑制效应可能还不够显著，但有一定的趋势（11 毫秒）。启动 1 和启动 2 条件下的效应很小。同时可以看到，各种条件下的标准差差异较大，表明数据可能不够稳定，通过增加被试有可能获得显著的启动或抑制效应。研究者继续增加被试至 75 人。从表 13-3 中可以看到，在 75 名被试的反应时平均数结果中，启动 3 条件下的抑制效应趋势加强了（20 毫秒），各种条件下的标准差差异减小了，表明数据更加趋于稳定。进一步的 F 检验结果表明，启动 3 条件下的抑制效应是统计上显著的。

表 13-3　75 名被试在启动实验 2 中反应时数据（单位：ms）

	被试数	平均数	标准差	启动效应
启动 1	75	751	107	−8
启动 2	75	743	104	0
启动 3	75	763	113	−20**
控制	75	743	103	

以上的例子告诉我们一组数据的平均数、标准差等描述统计信息的重要性。它可以帮助研究者了解初步结果的性质，确定进一步研究方向，或者考虑进一步的统计分析方法。

第二节　数据的转换

转换（transformation）是指对一组数据进行系统的转变。转换的过程中这组数据的某些特征改变了，而其他一些特征没有改变。在行为科学研究中，研究者有时需要在正式数据处理之前对原始数据进行转换，常见的原因可能包括：原始数据不能很好地满足 F 检验的需要，研究者希望获得误差变异的同质性，或者使原始数据接近正态分布，或者是希望减小误差变异等。在一些情况下，实验处理效应本身可能是存在的，但由于原始数据的变异非常大，误差变异过大而造成处理效应不显著。在这样的情况下，数据转换可能对提高实验的敏感性有帮助。还有一些情况下，当原始数据中有极端值，如在百分率数据中有大量的 1 和 0 的极端分数时，明显破坏了数据的正态分布，合适的数据转换会使数据趋于正态分布。一般来说，当处理水平的平均数和误差变异是百分率时，以及当误差变异的分布是同质时，可以使用数据转换。本节中，我们将介绍一些常用的转换方法。

一、几种数据转换的方法

（一）平方根转换

在某种数据中，处理水平的平均数和变异都是可表示为百分数的。当因变量是一个出现的可能性很小的小概率事件时，经常会出现符合泊松分布，如在一个简单的迷宫实验中，小白鼠在每个选择点上的错误数量。对这样的数据经常可以进行平方根转换（square-root transformation），转换后，数据通常可以趋于正态化并减小变异。

转换分数的公式：

$$Y'=\sqrt{Y}$$

其中 Y 是原始分数。如果 Y 小于 10，更合适的转换是：

$$Y' = \sqrt{Y + 0.5}$$

或

$$Y' = \sqrt{Y} + \sqrt{Y+1}$$

转换后分数Y'的平均数和变异不再是可表示为百分数的，数据的变异变得更加同质。这些转换分数比原始分数更适合进行方差分析。我们举一个例子来说明平方根转换。在一个儿童汉字命名实验中，要对二、四、六年级儿童命名错误中的声旁错误和类比错误的比例进行分析。其中声旁错误指儿童用声旁去命名一个不规则字，如将“歧”读做 zhi（支），类比错误指儿童用一个带同声旁的熟悉字去命名一个不规则字，如将“歧”读做 ji（技）。这两种错误反映儿童对声旁和整字的关系有一定的认识，可以称做语音有关的错误。因变量是儿童汉字命名中语音有关错误的数量。由于不同年级、不同能力儿童的读音错误数量的差异非常大，导致数据的变异很大，正态性不好。从平均数数据的描述统计结果（表13-4）看，在二年级中，无论是能力高、中、低的儿童都很少犯语音有关的错误，而且不同能力儿童之间没有很大差异。四年级儿童犯语音有关错误的数量大大增加，而且不同能力儿童之间有很大差异。六年级儿童犯语音有关错误的数量仍然很高，但不同能力儿童之间的差异很小。数据表明在每个年级中，能力高、中、低儿童的汉字读音中语音有关错误变化的趋势似乎是不一致的。如果要探讨年级和能力对儿童的语音有关错误的影响，可以对数据进行 3（年级）×3（能力）两因素实验的方差分析。结果发现年级和能力的交互作用是不显著的。

表 13-4　不同年级、不同能力儿童平均读音错误数的原始数据

能力	年级		
	二	四	六
低	1.6	5.9	14.4
中	1.8	10.4	14.8
高	3.9	12.6	15.7

研究者尝试对原始数据进行了平方根转换（见表 13-5），转换数据使用的公式如下：

$$Y'=\sqrt{Y+0.5}$$

表 13-5 中是转换前和转换后的数据，其中“年级”一栏的数字 1、2、3 为二、四、六年级的编码，“能力”一栏的数字 1、2、3 为能力低、中、高的编码，“原始”一栏的数字为儿童犯读音有关错误的数量，“平方根”一栏为对原始错误数进行平方根转换后的数字。

表 13-5　不同年级儿童读音错误数的原始数据和平方根转换数据

年级	能力	原始	$\sqrt{n+0.5}$	年级	能力	原始	$\sqrt{n+0.5}$	年级	能力	原始	$\sqrt{n+0.5}$
1	1	0	0.7	2	1	6	2.5	3	1	21	4.6
1	1	3	1.9	2	1	1	1.2	3	1	16	4.1
1	1	1	1.2	2	1	9	3.1	3	1	9	3.1
1	2	4	2.1	2	2	11	3.4	3	2	17	4.2
1	2	4	2.1	2	2	10	3.2	3	2	6	2.5
1	2	0	0.7	2	2	6	2.5	3	2	19	4.4
1	3	7	2.7	2	3	15	3.9	3	3	18	4.3
1	3	5	2.3	2	3	14	3.8	3	3	22	4.7
1	3	4	2.1	2	3	12	3.5	3	3	9	3.1
…	…	…	…	…	…	…	…	…	…	…	…

仔细观察表 13-5 中的数据可以看到，错误数的原始数据的全距 (range) 是非常大的，使用平方根转换后，转换数据的全距明显减小。我们再来看一看转换前后数据的描述统计的情况。从表 13-6 中可以看到，使用平方根转换后，描述统计表明，与原始数据的标准差相比，转换数据的标准差的差异明显减小。

表 13-6 不同年级、不同能力儿童读音错误数的原始和平方根转换数据的描述统计

年级	能力	原始数据		平方根转换数据		被试数
		平均数	标准差	平均数	标准差	
二年级	低	1.555 6	1.943 7	1.310 8	0.616 1	9
	中	1.800 0	1.612 5	1.416 4	0.561 1	15
	高	3.923 1	2.722 2	1.994 3	0.695 0	13
		2.486 5	2.340 7	1.593 7	0.677 4	37
四年级	低	5.923 1	5.057 4	2.319 9	1.062 1	13
	中	10.400 0	4.171 3	3.238 3	0.665 5	15
	高	12.583 3	3.260 2	3.591 5	0.448 5	12
		9.600 0	4.960 4	3.045 8	0.918 9	40
六年级	低	14.400 0	4.550 9	3.820 3	0.582 2	10
	中	14.812 5	4.861 0	3.854 4	0.697 5	16
	高	15.700 0	4.083 8	3.994 3	0.522 7	10
		14.944 4	4.471 8	3.883 8	0.609 1	36
总计	低	7.343 8	6.597 2	2.505 0	1.272 6	32
	中	9.130 4	6.635 3	2.858 5	1.221 4	46
	高	10.257 1	6.035 9	3.113 3	1.046 7	35
		8.973 5	6.486 9	2.837 3	1.198 3	113

我们分别用原始数据和转换数据进行了方差分析，结果见表 13-7。可以看到，如果使用原始数据进行方差分析时，年级与能力的交互作用是不显著的 [$F(4, 104)=1.939$，$p=0.109$]。然而，如果对数据平方根转换后，使用转换数据进行方差分析时，年级与能力的交互作用是显著的 [$F(4, 104)=2.533$，$p=0.045$]。

表 13-7　SPSS 方差分析输出结果

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Raw Number	3198.998	8	399.875	27.470	.000
	Sqrt Transformation	112.815	8	14.102	30.554	.000
Intercept	Raw Number	8838.328	1	8838.328	607.155	.000
	Sqrt Transformation	876.597	1	876.597	1899.292	.000
GRADE	Raw Number	2765.296	2	1382.648	94.982	.000
	Sqrt Transformation	96.172	2	48.086	104.186	.000
ABILITY	Raw Number	194.793	2	97.397	6.691	.002
	Sqrt Transformation	8.278	2	4.139	8.968	.000
GRADE * ABILITY	Raw Number	112.893	4	28.223	1.939	.109
	Sqrt Transformation	4.677	4	1.169	2.533	.045
Error	Raw Number	1513.923	104	14.557		
	Sqrt Transformation	48.000	104	.462		
Total	Raw Number	13812.000	113			
	Sqrt Transformation	1070.500	113			
Corrected Total	Raw Number	4712.920	112			
	Sqrt Transformation	160.815	112			

（二）log 转换

如果在某种数据中，处理水平的平均数和标准差是可表示为百分数的，log 转换（logarithmic transformation）可能也是一种合适的选择。

log 转换的转换分数公式是：

$$Y' = \log_{10} Y$$

其中，Y 是指原始分数。当有些原始分数的值很小或者是 0 的时候，可以使用另一个公式：

$$Y' = \log_{10}(Y+1)$$

有些研究发现，当因变量是反应时测量，数据呈现正偏态分布时，log 转换是很有用的，能使正偏态的反应时数据很好地正态化。我们举一个例子来说明 log 转换。在一个研究中，研究者用眼动技术探讨语音、字形在汉语阅读中的作用，即通过改变关键词的形、音特征，观察被试阅读中眼动指标的变化。实验中要求被试正常阅读计算机屏幕上呈现的短文，

例如："我刚从国外回到家的时候，家里的燃料是木炭，后来烧的是煤粉捏的大煤球，不但灰大烟大，而且碎纸、劈柴、洋火，样样不能离。"每个短文中有一个关键词，在这段短文中关键词是"烧"。研究者设计了四种实验条件：原词（如"烧"）、同音而形不相似的词（如"捎"）、形似而不同音的词（如"绕"）、不同音形也不相似的词（如"唱"）。因变量记录的是被试的眼动数据，包括首次注视时间、总注视时间、回扫次数等。由于原始数据的正态分布不够好，研究者对数据进行 log 转换。从"首次注视时间"（表 13-8）和"总注视时间"（表 13-9）的表中可以观察到，进行 log 转换后，数据的标准差明显减小了。

表 13-8　首次注视时间的原始数据和 log 转换数据的平均数和标准差（单位：ms）

关键词	原始数据	log 转换数据
原词	214(81)	2.30(0.16)
同音而形不相似的词	234(145)	2.31(0.23)
形似而不同音的词	224(93)	2.32(0.16)
无关控制词	220(87)	2.31(0.15)

表 13-9　总阅读时间的原始数据和 log 转换数据的平均数和标准差（单位：ms）

关键词	原始数据	log 转换数据
原词	293 (181)	2.40 (0.24)
同音而形不相似的词	332 (242)	2.43 (0.28)
形似而不同音的词	328 (276)	2.43 (0.25)
无关控制词	420 (327)	2.51 (0.30)

我们再来看一看 log 数据转换后数据正态分布的情况。图 13-1 中是首次注视时间的原始数据的频率分布，图 13-2 中是首次注视时间的 log 转换数据的频率分布，可以明显看出 log 转换数据的分布更加接近正态分布。

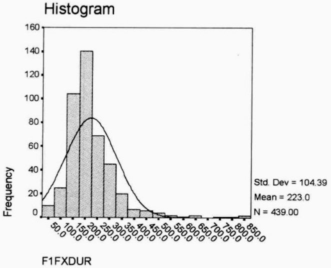


图 13-1　首次注视时间的原始数据的频率分布

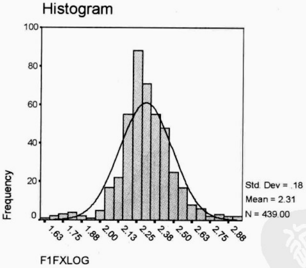


图 13-2　首次注视时间的 log 转换数据的频率分布

log 数据转换后，总阅读时间的数据正态分布情况有类似的变化。图 13-3 中是总阅读时间的原始数据的频率分布，图 13-4 中是总阅读时间的 log 转换数据的频率分布，同样可以看到 log 转换数据的分布更加接近正态分布。

Frequency Distribution for C：Total Duration

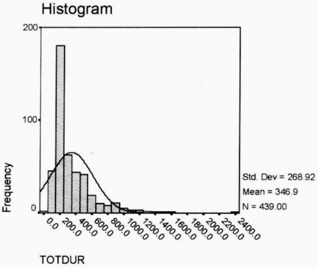


图 13-3　总阅读时间的原始数据的频率分布

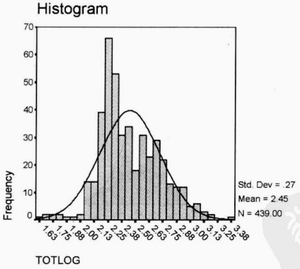


图 13-4　总阅读时间的 log 转换数据的频率分布

（三）角度转换

角度转换（angular transformation）一般用于当数据中各处理水平的平均数和变异是可表示为百分数的，同时数据分布中有极值的时候。例如，因变量测量是在不同处理中的正确反应数，当施测的数量固定时，由

于被试能力的差异，可能出现两项极值，即出现百分之百的正确率，或百分之零的正确率。在这种情况下，对数据进行角度转换是合适的。

角度转换分数的公式：

$$Y'=2\arcsin(Y^{1/2})$$

其中，Y 是指原始分数。当因变量测量分数被表示为百分数时，Y 的值一般是从 0.001 到 0.999。当数据中出现极值时，即 $Y=0$ 和 $Y=1$ 时，巴特利特（Bartlett）提出：当 $Y=0$ 时，可以用 $\frac{1}{2n}$ 或 $\frac{1}{4n}$ 来替换 0；当 $Y=1$ 时，可以用 $1-\frac{1}{2n}$ 或 $1-\frac{1}{4n}$ 替换 1。其中，n 是百分数所基于的观察的数量。我们再举一个例子对数据的角度转换进行说明。在一个儿童汉字命名实验中，要对二、四、六年级儿童对汉字的命名情况进行分析，因变量是命名正确率。由于不同年级、不同能力儿童的命名正确率差异非常大，导致原始数据中有 1 和 0 的极值出现，数据的正态性不好，研究者尝试对数据进行角度转换（Shu，Anderson & Wu，2000）。

转换数据使用的公式是：$Y'=2\arcsin(Y^{1/2})$。转换时对数据中一些 0 和 1 极值用 $\frac{1}{2n}$ 和 $1-\frac{1}{2n}$ 进行了替换。表 13-10 中是该实验的原始数据和角度转换数据。

表 13-10　不同年级儿童读音正确率的原始数据和角度转换数据

年级	能力	原始	角度	年级	能力	原始	角度	年级	能力	原始	角度
1	1	0.2	0.93	2	1	0.3	1.16	3	1	0.7	1.98
1	1	0.4	1.37	2	1	0.4	1.37	3	1	0.9	2.5
1	1	0.0	0.00	2	1	0.6	1.77	3	1	0.6	1.77
1	2	0.5	1.57	2	2	0.7	1.98	3	2	0.9	2.5
1	2	0.2	0.93	2	2	0.6	1.77	3	2	0.7	1.98
1	2	0.9	2.50	2	2	1	3.14	3	2	0.8	2.21
1	3	1.0	3.14	2	3	0.9	2.5	3	3	0.9	2.5
1	3	0.9	2.50	2	3	1	3.14	3	3	1	3.14
1	3	0.6	1.77	2	3	0.8	2.21	3	3	0.9	2.5
…	…	…	…	…	…	…	…	…	…	…	…

表13-10中是转换前和转换后的数据，其中“年级”一栏的数字1、2、3为二、四、六年级的编码，“能力”一栏的数字1、2、3为能力低、中、高的编码，“原始”一栏的数字为儿童的汉字命名正确率，“角度”一栏为对原始正确率进行角度转换后的数字。表13-11中是转换前和转换后对数据进行描述统计分析的结果。

表13-11 不同年级、不同能力儿童读音正确率的原始和角度转换数据的描述统计

年级	能力	原始数据		角度转换数据		被试数
		平均数	标准差	平均数	标准差	
二年级	1	0.333 3	0.250 0	1.102 1	0.713 9	9
	2	0.466 7	0.303 9	1.444 1	0.784 6	15
	3	0.915 4	0.121 4	2.745 1	0.482 5	13
	总计	0.591 9	0.340 2	1.818 0	0.962 8	37
四年级	1	0.253 8	0.194 1	0.958 5	0.551 7	13
	2	0.500 0	0.210 4	1.598 2	0.549 3	15
	3	0.691 7	0.210 9	2.068 2	0.605 0	12
	总计	0.477 5	0.266 5	1.531 3	0.710 6	40
六年级	1	0.530 0	0.254 1	1.635 1	0.562 8	10
	2	0.693 8	0.194 8	2.002 4	0.434 7	16
	3	0.860 0	0.135 0	2.510 2	0.493 8	10
	总计	0.694 4	0.230 5	2.041 4	0.579 5	36
总计	1	0.362 5	0.252 4	1.210 3	0.655 4	32
	2	0.556 5	0.256 2	1.688 5	0.637 2	46
	3	0.822 9	0.184 8	2.445 9	0.592 4	35
	总计	0.584 1	0.294 2	1.787 7	0.789 4	113

从图13-5和图13-6中可以看到，对数据进行了角度转换后，数据的正态分布更好，更加适于进行方差分析。

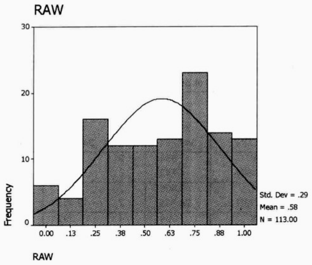


图 13-5　读音正确率的原始数据的频率分布

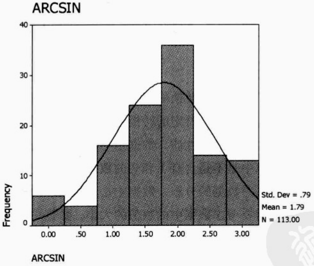


图 13-6　读音正确率的角度转换数据的频率分布

分别用原始数据和角度转换数据进行了方差分析，结果见表 13-12。结果表明，如果使用原始数据进行方差分析时，年级与能力的交互作用是不显著的［$F(4, 104)=1.978$，$p=0.103$］。然而，如果对数据角度转换后，使用转换数据进行方差分析时，年级与能力的交互作用是显著的

[$F(4, 104)=2.496$, $p=0.047$]。

表 13-12 SPSS 输出的方差分析结果

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	RAW	4.845	8	.606	12.998	.000
	ARCSIN	34.528	8	4.316	12.726	.000
Intercept	RAW	36.965	1	36.965	793.282	.000
	ARCSIN	346.784	1	346.784	1022.509	.000
GRADE	RAW	.834	2	.417	8.950	.000
	ARCSIN	4.742	2	2.371	6.991	.001
ABILITY	RAW	3.406	2	1.703	36.548	.000
	ARCSIN	24.935	2	12.468	36.762	.000
GRADE * ABILITY	RAW	.369	4	.092	1.978	.103
	ARCSIN	3.386	4	.847	2.496	.047
Error	RAW	4.846	104	.047		
	ARCSIN	35.272	104	.339		
Total	RAW	48.240	113			
	ARCSIN	430.934	113			
Corrected Total	RAW	9.691	112			
	ARCSIN	69.800	112			

二、转换方法的选择

我们已经介绍了几种数据转换的方法。从以上描述中可以看到，在某种特定的情景下，对数据进行特定的转换是更有效的。因此要选择合适的转换方法，首先要考虑数据的性质。数据是不是可表示为百分数的？在可表示为百分数的数据中是不是含有 1 和 0 的数据？数据记录的是错误数还是反应时？其次要考虑数据的分布。数据中是不是有两项极值？是不是正偏态分布的？最后需要考虑数据中的数值的大小。数据中是否含有大量的 0 或接近 0 的数据？考虑以上描述的规则有助于研究者选择合适的数据转换方式。但是需要注意的是，并不一定对任何一组数据总是能找到一种合适的转换方法。

选择合适的数据转换方式的另一种途径是尝试使用每一种转换得到每一个处理水平的最大分数和最小分数，然后确定每个水平内数据的全距，计算最大值与最小值的比率，产生最小比率的转换是可以考虑选择的合适的转换。这种选择方法主要适用于当研究者关心如何减小误差变异的

时候。

我们通过举例来说明选择转换方法的过程。如果我们有一组数据，希望找出对该组数据合适的转换方法。原始数据中有三个处理水平，每一个水平有五名被试接受了实验处理。

表 13-13 三种处理条件下的原始数据

A1	A2	A3
4	7	13
1	5	7
5	3	8
4	6	11
3	8	7

可以看到，这组数据的原始数据变异是比较大的，如果想要找到合适的转换方法，以减小变异，可以进行下列的尝试。第一步，先找到每一个处理水平的最大分数和最小分数。在这组原始数据中，三个处理水平的最大分数分别是5(A1)、8(A2)、13(A3)，三个处理水平的最小分数分别是1(A1)、3(A2)、7(A3)。第二步，确定每个处理水平内的全距，分别为5－1＝4(A1)，8－3＝5(A2)，13－7＝6(A3)。第三步，计算出全距最大值与最小值的比率，在原始数据中的比率为6÷4＝1.5（见表13-12）。第四步，我们分别计算平方根转换分数和log转换分数，找到每一个处理水平的最大分数和最小分数，计算出最大值与最小值的比率。从表13-12中可以看到，平方根转换后，每个处理水平内的全距分别为2.35－1.22＝1.13(A1)，2.92－1.87＝1.05(A2)，3.67－2.74＝0.93(A3)，全距最大值与最小值的比率为1.13÷0.93＝1.22。log转换后，每个处理水平内的全距分别为0.78－0.30＝0.48(A1)，0.95－0.60＝0.35(A2)，1.15－0.90＝0.25(A3)，全距最大值与最小值的比率为0.48÷0.25＝1.92。比较原始分数、平方根转换分数和log转换分数的全距比率，得到比率最小的转换方法是最合适的转换。从表13-14中可以看出，对该组数据进行平方根转换可能是相对更合适的。

表 13-14 选择原始数据转换方法的比较

	处理水平			$\frac{全距_{最大}}{全距_{最小}}$
	A1	A2	A3	
原始分数				
最大分数（L）	5	8	13	
最小分数（S）	1	3	7	$\frac{6}{4}=1.5$
全距	4	5	6	
平方根转换				
$\sqrt{L+0.5}$	2.35	2.92	3.67	
$\sqrt{S+0.5}$	1.22	1.87	2.74	$\frac{1.13}{0.93}=1.22$
全距	1.13	1.05	0.93	
log 转换				
$\log(L+1)$	0.78	0.95	1.15	
$\log(S+1)$	0.30	0.60	0.90	$\frac{0.48}{0.25}=1.92$
全距	0.48	0.35	0.25	

为什么我们可以对原始数据进行数据转换？我们知道，心理学实验研究中选择什么指标做因变量常常是研究者任意确定的。不同的实验可能选择不同的因变量指标，有时在同样的实验设计中选择不同的因变量指标都是可能的。例如，一个关于句子加工过程中词汇语义违反的研究，研究者可能用反应时做因变量指标，也可能使用眼动的首次注视时间、回扫次数等做因变量指标。如果数据呈正偏态分布，或者数据变异较大导致方差分析不显著是与人为选择测量量表的类型有关的话，那么数据的变异较大、正偏态分布等状况是可以通过数据转换或改变测量量表得到改善的。合适的数据转换会使数据更适合于进行方差分析，以增加 F 检验显著的可能性。然而，通过数据转换增加 F 检验显著的可能性只适合于当实验的处理效应本身是存在的情况。如果处理效应本身是不存在的，任何种类的数据转换都是没有作用的。另外，获得显著的处理效应主要是依靠好的实验设计和实施，通过数据转换来改善方差分析结果的可能性是非常有限的。

一旦对一组数据选择了合适的转换方法，数据就会在一个新的量表下进行分析，所有有关处理效应的推论是在新量表的角度下产生的。在多数行为研究情景下，基于 log 转换的分数、基于平方根转换的分数所作的推论和基于未经转换的分数所作的推论的意义是不变的。

第三节　不等组实验数据的分析

在方差分析实验中，研究者一般会对不同处理条件下的被试数量进行控制，确保在各个处理条件下的被试数量相等，但有时也会出现在不同的处理条件下被试数量不相等的情况。这种情况的产生可能有多种原因，一个常见的原因是由于实验中不同的处理条件下使用自然的、固定团体，如使用在班级、车间等固定团体中的被试，而这些固定团体中的被试数量可能是不相等的。例如，在比较不同教学方法的研究中，很难做到将学生随机分配进各种教学条件，因此研究者常常保持在原有的班级中进行各种教学或测验，但每个班级的学生数量可能是不相等的。另外，在一些固定团体的研究、长期追踪研究中可能由于生病、转学或不能完成测验等与任务无关的原因导致被试的缺失，从而造成各种条件下被试不相等的情况。再有，有时在实验完成后可能对数据进行再分析，根据实验中的新发现对处理水平进行重新定义，而不是使用原实验设计中的处理水平分组，这时也会出现不等组的情况。我们在本节中重点讨论由于各种情况导致的各个实验处理条件下被试数量不相等时的数据分析，这与第六章中介绍的不等组设计的概念是不同的。

我们应当如何处理不等组的实验数据？现实中经常有两种选择。

（1）在每种条件下选取同样数量的被试或同样数量的实验材料进行数据分析。但选取哪些被试、哪些材料参加数据分析，根据什么原则保留或舍弃部分被试或数据，这在如何操作上存在问题。

（2）使用合适的统计处理。例如，使用不等组的数据处理，可以避开数据取舍的问题。

本节中将介绍处理不等组数据的两种统计方法：无权重的分析（unwighted）和权重的分析（wighted）。使用这两种统计方法处理不等组

数据时不需要去除数据，以便达到等组的目的。但是，进行不同处理条件下的不等组被试的数据分析有一些重要的前提。例如，各种处理条件下被试数量不相等的原因应与分配实验处理条件无关。在一般的实验设计中，我们强调随机分配被试进入各个实验处理条件。随机分配被试的好处是保证在每一个实验处理条件下被试数量是相等的，被试是同质的。然而，在不等组情况下这种好处可能被削弱或消失，实验结果的科学性可能会受到影响。

一、单因素实验中的不等组数据计算

无权重的分析和权重的分析是两种处理不等组数据时使用的统计方法。使用无权重或权重方法分析不相等被试组数据的唯一差别是平方和 SS 的计算。在无权重的分析中，每个处理平均数对确定平方和 SS 有同等的贡献。而在有权重的分析中，每个处理平均数对确定平方和 SS 的贡献与每个组被试的数量有关。或者说，每个处理平均数对 SS_A 的影响是由在某种条件下被试数量的比例所决定的。这样，有权重的分析中，数量大的被试组在分析中会比数量小的被试组计数得更多，贡献更大。而在无权重分析中，数量大的被试组和数量小的被试组的贡献是相同的。这个方法适用于每个组（单元）中含有至少一个被试。使用无权重或权重分析的前提是假定不等组的出现是与实验处理条件分配无关的。

下面我们将介绍用权重分析或无权重分析不相等被试组的两种公式。首先我们看一下当被试组相等时的处理平方和的计算公式。根据公式

$$SS = \sum_{i=1}^{n}(X_i - \overline{X})^2$$

被试组相等时使用的公式如下：

$$SS_A = n\sum_{j=1}^{p}(\overline{A_j} - \overline{T})^2$$

其中 $\overline{A_j}$ 表示每个实验处理组的平均数，$\overline{T}$ 表示总平均数，p 表示自变量的水平数，n 表示每组被试的数量。

当被试组是不相等的时候，使用权重分析的处理平方和的计算公式如下：

$$SS_A = \sum_{j=1}^{p} n_j(\overline{A_j} - \overline{T})^2$$

其中 n_j 表示各不等组的被试数量。

比较以上两个公式，可以看到主要变化是被试数 n 的计算不同。在等组数据的分析中，可以使用以上任意一个公式计算，因为两种方法计算的结果是相同的。而在不等组数据的分析中，只能用第二个公式，因为只有第二个公式中有合适的方式表示各组的被试数量 n_j。

可以看到，无权重分析的公式和相等被试组的公式是相同的。下面我们举两个例子来说明当被试是不等组时，权重分析和无权重分析两种计算方法的异同。

在一个单因素完全随机实验中（见表 13-15，例 1），自变量有 A1、A2、A3 三个水平，由于某种特殊的与实验处理条件分配无关的原因，每种处理水平下的被试数是不相等的：第一组 15 人，第二组 10 人，第三组 25 人。三个组的因变量原始总分数分别为 45、10 和 100，三组平均数分别为 3、1 和 4。研究者要探讨的问题是：三组平均数存在显著差异吗？

下面我们分别用权重分析和无权重分析方法进行计算，并观察两种计算方法对结果的影响。

表 13-15　两个实验例子的处理平均数数据

	例 1			例 2		
	A1	A2	A3	A1	A2	A3
总分数	45	10	100	30	25	60
被试数	15	10	25	10	25	15
平均数	3	1	4	3	1	4

首先，我们用权重分析的方法计算处理效应 SS_A。

第一步，用加权平均数（weighted mean）方法，进行总平均 $\overline{T}$ 的计算：

$$\overline{T} = \frac{(3\times15)+(1\times10)+(4\times25)}{15+10+25} = \frac{155}{50} = 3.1$$

第二步，进行 SS_A 的计算：

$$SS_A = \sum_{j=1}^{p} n_j(\overline{A_j} - \overline{T})^2$$
$$= 15 \times (3 - 3.1)^2 + 10 \times (1 - 3.1)^2 + 25 \times (4 - 3.1)^2$$
$$= 64.5$$

然后，我们再用无权重分析的方法计算处理效应 SS_A。

第一步，用调和平均数（harmonic mean）方法，进行 n 的计算：

$$n = \frac{1}{\frac{1}{3} \times \left(\frac{1}{15} + \frac{1}{10} + \frac{1}{25}\right)} = 14.52$$

第二步，进行总平均 $\overline{T}$ 的计算：

$$\overline{T} = \frac{3+1+4}{3} \approx 2.67$$

第三步，进行 SS_A 的计算：

$$SS_A = n\sum_{j=1}^{p} (\overline{A_j} - \overline{T})^2$$
$$= 14.52 \times [(3 - 2.67)^2 + (1 - 2.67)^2 + (4 - 2.67)^2]$$
$$= 67.81$$

结果可以看到，两种方法计算的处理效应 SS_A 是不相同的，主要原因是与被试不等组有关。关于加权平均数和调和平均数计算的细节，请见《现代心理与教育统计学》(张厚粲和徐建平，2004)。

那么，哪种计算受不等组的影响更大呢？不等组问题是如何影响处理效应的计算呢？我们通过另外一组数据（例 2）的计算来观察权重分析和无权重分析对计算结果的影响。

我们有另外一组数据，也是一个单因素完全随机实验（见表 13-15，例 2），自变量有三个水平，每种处理水平下的被试数也是不相等的：第一组 10 人，第二组 25 人，第三组 15 人。三个组的因变量原始总分数分别为 30，25 和 60，三个组的平均数同样是 3、1 和 4。研究者感兴趣的是同样的问题：这三组平均数存在显著差异吗？

我们同样用权重分析和无权重分析方法进行计算，并观察两种计算方法对结果的影响。

首先用权重分析方法计算 SS_A。

总平均 $\overline{T}$ 的计算：

$$\overline{T}=\frac{(3\times10)+(1\times25)+(4\times15)}{10+25+15}=2.3$$

SS_A 的计算：

$$\begin{aligned}SS_A &= \sum_{j=1}^{p} n_j(\overline{A_j}-\overline{T})^2\\ &= 10\times(3-2.30)^2+25\times(1-2.30)^2+15\times(4-2.30)^2\\ &= 42.25\end{aligned}$$

然后用无权重分析方法计算 SS_A。

n 的计算：

$$n=\frac{1}{\frac{1}{3}\times\left(\frac{1}{10}+\frac{1}{25}+\frac{1}{15}\right)}=14.52$$

总平均 $\overline{T}$ 的计算：

$$\overline{T}=\frac{3+1+4}{3}\approx2.67$$

SS_A 的计算：

$$\begin{aligned}SS_A &= n\sum_{j=1}^{p}(\overline{A_j}-\overline{T})^2\\ &= 14.52\times[(3-2.67)^2+(1-2.67)^2+(4-2.67)^2]\\ &= 67.81\end{aligned}$$

从计算结果中我们再次看到，两种方法计算的处理效应 SS_A 是不相同的。如果我们比较用权重和无权重两种方法计算例 1 和例 2 的数据可以发现，在两个实验例子中总样本数是相同的，三种处理条件的平均数是相同的。但是在两个举例中分配给三种处理条件的三组被试数（n_1，n_2，n_3）变化了，与各组被试相对应的总分数变化了。这些变化在无权重分析两组数据时没有影响：在两个例子的计算中，$n=14.52$ 和 $\overline{T}\approx2.67$ 都是不变的，处理效应（$SS_A=67.81$）也是相同的。而在权重分析中，两个例子中由于在三种处理条件下被试数变化以及相应总分数的变化，计算得到不同的结果，分别得到 $\overline{T}=3.1$（例 1）和 $\overline{T}=2.3$（例 2），从而导致例 1（$SS_A=64.5$）和例 2（$SS_A=42.25$）的处理效应有很大差异。从两个结果的比较中可以看到，无权重分析是不受取样大小、分布影响的，而权

重分析是受取样大小、分布的影响的。或者说无权重分析假设，虽然各组被试数不相等，但每个组对处理效应的贡献是相同的。而权重分析假设，每个组对处理效应的贡献是受各组被试数量影响的。

二、两因素实验中的不等组数据计算

我们再举一个两因素实验的例子（例 3）。实验中有 A 和 B 两个因素，每个因素有两个水平。在四个处理结合上被试的数量不同，分别为 3、3、3、2。被试的不等组会对处理效应的计算产生影响吗？当在两因素实验中有不等被试组时，首先要确定进行无权重还是有权重的平均数分析。这时需要考虑的关键问题是：我们是希望处理效应受各组不同样本大小的影响还是希望每个组对处理效应的贡献是相同的，即处理效应不受各组样本大小的影响？通常情况下，人们会选择无权重分析，因为通常我们的实验检验假说中并不关心不等取样组的效应。下面我们主要介绍两因素实验中有不等被试组时，无权重的平均数分析过程。该实验的数据如下（见表 13-16）。

表 13-16 两因素实验举例（例 3）的原始数据

	A1B1	A1B2	A2B1	A2B2
	9	6	5	4
	8	4	4	2
	7	2	3	
总分数	24	12	12	6
被试数	3	3	3	2

首先，我们要计算平均数数据。

表 13-17 两因素实验无权重分析的平均数数据

	A1	A2	Σ
B1	8	4	6
B2	4	3	3.5
Σ	6	3.5	4.75

然后，计算我们再使用无权重分析计算 SS_A、SS_B 和 SS_{AB}。

$$\overline{T}=\frac{8+4+4+3}{4}=\frac{19}{4}=4.75$$

$$n=\frac{1}{\frac{1}{4}\times\left(\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{2}\right)}=2.67$$

$$\begin{aligned}SS_A &= n\sum_{j=1}^{p}(\overline{A_j}-\overline{T})^2 = 2.67\times2\times[(6-4.75)^2+(3.5-4.75)^2]\\ &= 2.67\times2\times(1.56+1.56)\\ &= 16.66\end{aligned}$$

$$\begin{aligned}SS_B &= n\sum_{k=1}^{q}(\overline{B_k}-\overline{T})^2 = 2.67\times2\times[(6-4.75)^2+(3.5-4.75)^2]\\ &= 2.67\times2\times(1.56+1.56)\\ &= 16.66\end{aligned}$$

$$\begin{aligned}SS_{AB} &= n\sum_{j=1}^{p}\sum_{k=1}^{q}(\overline{AB_{(jk)}}-\overline{T})^2-SS_A-SS_B\\ &= 2.67\times[(8-4.75)^2+(4-4.75)^2+(4-4.75)^2+(3-4.75)^2]-\\ &\quad SS_A-SS_B\\ &= 2.67\times(10.56+0.56+0.56+3.06)-SS_A-SS_B\\ &= 2.67\times14.74-SS_A-SS_B\\ &= 39.35-16.66-16.66\\ &= 5.93\end{aligned}$$

其中$\overline{A_j}$表示 A 因素每个处理组的平均数，$\overline{B_k}$表示 B 因素每个处理组的平均数，$\overline{T}$ 表示总平均数，p 表示 A 因素的水平数，q 表示 B 因素的水平数，n 表示各不等组被试的平均数量。

表 13-18 两因素实验无权重分析的方差分析表

变异来源	SS	df	MS	F	p
A	16.66	1	16.66	8.33	0.023
B	16.66	1	16.66	8.33	0.023
AB	5.93	1	5.93	2.97	0.127
误差	14	7	2.00		
合计	53.25	10			

从方差分析表中可以看到，A 因素和 B 因素的主效应都是显著的。

最后，我们使用 SPSS 软件计算例 3 的数据。描述统计（见表 13-19）的结果输出如下：

表 13-19　SPSS 输出的描述统计结果

A	B	Mean	Std. Deviation	N
1.00	1.00	8.0000	1.0000	3
	2.00	4.0000	2.0000	3
	Total	6.0000	2.6077	6
2.00	1.00	4.0000	1.0000	3
	2.00	3.0000	1.4142	2
	Total	3.6000	1.1402	5
Total	1.00	6.0000	2.3664	6
	2.00	3.6000	1.6733	5
	Total	4.9091	2.3433	11

例 3 的数据的方差分析结果（表 13-20）输出如下：

表 13-20　SPSS 输出的方差分析结果

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	40.909	3	13.636	6.818	.017
Intercept	240.667	1	240.667	120.333	.000
A	16.667	1	16.667	8.333	.023
B	16.667	1	16.667	8.333	.023
A * B	6.000	1	6.000	3.000	.127
Error	14.000	7	2.000		
Total	320.000	11			
Corrected Total	54.909	10			

将手工计算和计算机计算的结果进行比较，可以看出，SPSS 计算使用的是无权重分析方法。这种方法较少受各组被试数不相等的影响。

三、不等组数据对 *F* 值的影响

我们从上面举例的结果和分析中看到，当实验数据是不等组的时候，随着实验中各种处理条件下被试数目的变化，权重分析方法计算的处理平方和会产生变化。然而，无权重分析方法的处理平方和较少受各组被试数量不相等的影响。由于我们对各种处理条件下被试数量变化对处理效应的

影响不感兴趣，相比之下，无权重分析方法计算有较多的优势，因此更多的研究者处理数据时使用无权重分析方法。但是，不等组被试问题真的不会对处理效应带来影响，以至于我们可以忽略不计吗？需要指出的是，实际上不等组被试的数据最终是会对处理效应带来影响的，其影响主要表现在误差变异方面。我们可以举一个例子说明它的影响。表 13-21 中的几组数据分析是来自等组或不等组被试的实验。例如在实验 1 中，被试是等组的，每个处理条件下三个被试。在实验 2、实验 3、实验 4 中，被试都是不等组的。尤其在实验 4 中，四个处理条件下的被试数量相差很远，分别为 1、1、1 和 9 个被试。从表中可以看到，随着各个处理条件下被试数量差异的增大，无权重分析方法计算的 n 减小。即使当数据中保持处理效应 MS_A 不变，由于 n 的数值减小，引起误差均方 MS_E 增大，F 值减小，最终导致 p 值的显著性降低。

表 13-21　不等组数据引起的 F 值的变化

实验	各组被试数	N	MS_A	MS_E	F（3，8）	p
1	3，3，3，3	3.00	9.00	0.833	10.80	0.012
2	2，2，4，4	2.67	9.00	0.938	9.59	0.015
3	1，1，5，5	1.67	9.00	1.500	6.00	0.040
4	1，1，1，9	1.29	9.00	1.944	4.63	0.064

从以上的举例分析中我们可以明确地看到，不等组被试问题会对处理效应带来影响，以至对我们的实验结论带来影响。因此，我们在实验中应当尽可能避免被试不等组问题。

第四节　统计检验力

进行数据处理重要的目的是检验假说，获得显著性的证据。一个固定效应模型的 F 检验的检验力（power）是指拒绝虚无假说的可能性。检验力的知识对估计一个实验的区分度、确定使用样本的大小是非常有用的。

一、实验的敏感性与误差变异

实验的敏感性对研究者是非常重要的。如果一个实验中研究者期望的

处理效应不显著时，其原因可能是：（1）处理效应确实不存在；（2）实验设计不够敏感，无法检验出处理效应。我们曾经在“方差分析概论”一章说过，完全随机实验设计中，误差变异的估计是基于接受相同实验处理的被试之间存在的差异，然后把不同被试组的误差变异合并形成组内平方和。可以说，组内平方和是“纯粹”的对误差变异的估计。然而，组间平方和则不是一个对处理效应的“纯粹”的估计。由于实验处理是实施于不同的被试组，组间平方和实际不仅包含了处理带来的效应，同时也包含了无法区分出来的误差变异。实验的处理效应、主效应或交互作用是否显著是通过 F 值计算的。F 值的分子是处理效应，其中包含了无法区分出来的误差变异，而分母是误差变异。

$$F=\frac{\text{处理效应}}{\text{误差变异}}$$

一个实验的敏感性与误差变异的大小是什么关系？从 F 值的计算可以看出，对于一定量的处理效应，误差变异增加将导致 F 值的减小，意味着减小了拒绝虚无假说的概率。相反，减小误差变异将导致 F 值的增大，意味着增加获得显著性的概率。因此，可以说误差变异的减小可以增加实验的敏感性。在心理学研究中，误差有三种主要来源：随机变异、未控制因素的变异或无关变异、被试个体误差。这些误差都会反映在被试的因变量分数中。下面我们将细致分析这三种误差变异。

（一）随机变异

随机变异主要指实验中的一些随机误差。很多因素会影响被试的因变量分数，如设备的稳定性、主试的指导语和操作、环境因素（如照明、噪声、温度等）都可能有变化。这些因素在某种程度上都会影响实验中测量的行为，导致被试与被试之间的变异，影响实验误差的估计。我们通常会发现，即使在同样的实验处理条件下，每个被试的因变量分数也不是完全相同的，误差变异的来源之一是以上所说的这些随机变异。减小误差的方法包括小心调整设备、训练主试、使用专门的实验室等。如果某些因素难以控制，或者我们对这些因素特别感兴趣，我们也可以将其放入自变量，以便研究它们的影响。

（二）未控制因素的变异或无关变异

未控制因素的变异或无关变异主要指实验中有些因素会影响因变量分

数，但这些因素没有在实验设计时被包括在自变量中。例如，在一个记忆实验中，研究者在每天不同的时间（上午、下午、晚上）测试不同的被试，测试时间的不同可能会影响被试的行为。但研究者对测试时间对被试行为的影响并不感兴趣，因而测试时间并没有被包含在自变量之内。这时，测试时间就成为一个无关变量。无关变量也会导致被试与被试之间的变异，影响对实验误差的估计。减小误差的方法包括使用实验设计的方法在这些无关变量上进行匹配，如随机区组设计、拉丁方设计等，以便分离出无关变异的影响。也可以将无关变量中的某些因素放入自变量，以便研究它们的影响。

（三）被试个体误差

行为科学研究中误差变异的主要来源是被试的个体误差。被试的许多特点，如性别、受教育程度、气质、性格、态度等，都可能使得被试在完成实验任务的行为上是不同的。当将他们被随机分配到不同的处理条件，这种被试之间的差异就成为重要的误差变异的来源。可以较好地减小被试变异的方法包括：将被试分为区组或匹配被试、使用重复测量实验设计、进行协方差分析等。

将被试分为“同质”组可以减小实验误差，其原理是因为同质组中被试的变异会小于不同质组的被试变异。基本的方法是：对被试进行前测，或利用已有的测量，将被试分为区组，使区组内的被试尽量在一个或多个与因变量测量有关的特征上相似，而区组间的差异尽可能大。每个区组中的被试随机分配给不同的实验处理条件。这种设计通过分离区组变异、限定组内变异，达到减小误差变异的目的。

重复测量设计与随机区组设计的思想是相似的，由于一个被试接受所有的实验处理，因此匹配更加完善，更好地减小了被试变异。一般来说，在重复测量的混合实验的方差分析中，与被试内因素有关的误差变异一定会小于与被试间因素有关的误差变异，误差减小的原因是由于在前者中，被试个体误差被从误差变异的估计中分离了出去，而后者中还保留着被试的个体误差。当然，重复测量设计会导致学习、记忆等问题变得不可忽视，不是在任何情况下都可以使用。

协方差分析主要是提供一种对实验结果的调整，这种实验中被试之间

事先存在的差异可能会影响因变量的观测值，因此通过调整对误差的估计和处理效应的估计，可以减少误差变异。协方差分析对实验结果调整程度的大小依赖于选择的协变量与因变量之间的相关。

二、检验力和 F 检验

（一） Ⅰ型和Ⅱ型错误的关系

一个实验的敏感性的量化指标是它的检验力。检验力指当备择假说为真时，拒绝虚无假说的可能性。或者说，检验力表示对虚无假说的统计检验得出“现象存在”结论的可能性。也可以说，检验力是当虚无假说为假，作出正确决策的可能性。

我们进行的方差分析中，要估计的统计假说是处理效应等于零（虚无假说）。我们决定拒绝或者不拒绝虚无假说是基于 F 值。而 F 值可能导致错误的推论。如果没有处理效应，我们拒绝了虚无假说，就犯了Ⅰ型错误（也叫 α 错误）。如果处理效应存在（虚无假说为假），而我们不能拒绝虚无假说，就犯了Ⅱ型错误（也叫 β 错误）。在行为科学研究中，Ⅱ型错误表示对处理效应的视而不见的错误，它否定了一个事实上存在的效应或关系。

检验力可表示为：

$$检验力=1-\beta$$

检验力通常定义为当虚无假说为假时，拒绝虚无假说的可能性。随着Ⅱ型错误的可能性减小，检验力增加，测验的敏感性提高。换一种说法，检验力是不犯Ⅱ型错误的可能性。即当处理效应存在时，没有忽略它的可能性。

（二） Ⅰ型和Ⅱ型错误的控制

我们通常将 α 水平设置成一个固定值，通过选择 F 分布的拒绝区（α 水平），可以控制Ⅰ型错误的大小。然而，Ⅱ型错误（检验力）大小的控制通常不是这样直接的。一般研究者不是通过增大 F 分布的拒绝区，以便减小Ⅱ型错误，而是在设置固定的 α 水平后，通过其他途径增加实验的检验力。通常考虑的因素和使用的方法如下。

（1）取样大小：每种实验条件下观察的数量增加，检验力增加。

(2) 未控制的变量或无关变异：当分离出未控制的变量或无关变异源，检验力增加。

(3) 处理效应：处理效应的大小增加，检验力增加。

确定一个实验的检验力水平是很有用的。如果在一个实验里，我们已经估计了变异的来源，并且不能拒绝虚无假说，在放弃这个实验之前，我们应当估计实验的敏感性，以察觉处理效应实际存在的可能。我们可能假设实验中得到的差异是真正的差异，但由于处理效应本身很小，我们目前的实验设计或方法可能不能拒绝虚无假说。在这种情况下，虽然我们的实验中不能肯定处理效应的存在，但是我们可能也没有完全放弃备择假说，因为目前的结果还可能是由于实验设计对处理间差异相对不敏感造成的。一个后期检验力的确定可以引导研究者作进一步的研究选择：放弃现在的研究，或选用更敏感的实验。

在计划实验和解释统计不显著的结果的时候，对检验力的理解是非常重要的，它可以帮助研究者进一步分析结果不显著的原因和如何改进实验设计。但是通常的困难在于如何找到一种途径去确定检验力的值，更重要的问题是如何确认什么是合理的备择假说。在方差分析中，实验实施前确认合理的备择假说等于能够事先说出要进行的研究中处理效应的大小，但大多数研究者在实施实验之前是无法作出对检验力的估计的。

假定我们能够作出一个研究中处理效应大小的可靠估计，这对我们有什么好处呢？如果我们可以事先估计出研究中处理效应的大小，我们就可以检验自己实验的敏感性，确定可接受的处理效应的最小值。例如，我们可以比较估计的最小处理效应值和研究中实际得到的处理效应的实际值，从而确定处理效应是否真正存在。另外，我们也可以估计实验的误差变异，从而确定 F 值的不显著是否来源于实验的不敏感。再者，我们还可以通过预实验估计我们的实验所需要的检验力。

在一个实验中，我们总是期望增加实验的敏感性或检验力。但是达到多大的检验力是合适的？根据Ⅰ型错误和Ⅱ型错误的关系，由于研究者一般保持 α 在 0.05 置信区间水平，检验力的增加有一个实际的限定。我们需要使用其他方法增加检验力，如增加取样的数量、减小无关的变异、增加期望的处理效应等，其中最容易操作的是增加被试的取样数量。

（三）取样大小的确定

假定在一个实验研究中，我们可以估计处理效应的大小、误差变异的大小以及我们要达到的检验力，那么我们就可以确定为达到实验预期的敏感度所需要的取样大小。检验力分析对研究者的意义在于：在研究的计划阶段，它可以帮助确定要达到预定的可能存在的处理效应值所需要的被试数量；在研究的结果分析阶段，它对已完成的研究进行估计，确定在预定的 α 水平上没有发现显著的处理效应是否可能主要是由于被试的数量不够。

我们经常发现，在不同内容、不同方法的实验中，研究者使用的被试数量是不同的。研究者在实验设计时经常将检验力问题包含在其考虑中，特别是对有经验的研究者，当他们在某个领域进行了一段研究后，他们对实验设计的一般敏感性有一个“直觉”。一般的原则是，在处理效应较大的实验（如正常组与障碍组比较实验）或误差变异较小的实验（如字词启动实验）中可以使用较少的被试，在处理效应较小或误差变异较大的实验（如句子加工实验、眼动实验）中使用较多的被试。例如，在一般的成人的字词启动反应时实验中，通常处理的效应比较小，但被试的误差变异也比较小。每个版本中使用 20～30 个被试，即可以得到显著的启动效应。下面，我们举几个例子来说明被试取样大小对实验结果的影响。

通常不同年龄儿童行为实验的处理效应较大，但由于儿童个体差异也较大，要得到稳定可靠的效应，仍需要较多被试。例如，在一个考察儿童学习和记忆不同类型形声字读音的实验中，研究者对儿童学习和记忆的正确率进行了分析（见表 13-22）。其中，类型 1 为规则—一致字（即声旁读音与整字读音完全相同，且含有该声旁的所有字的读音都相同），类型 2 为规则—不一致字（即声旁读音与整字读音完全相同，但含有该声旁的所有字的读音不完全相同），类型 3 为声旁不知字（即声旁是儿童不熟悉、不知道读音的）。

从表 13-22 可以看到，当被试取样大小不同时，统计结果也不同，这将直接影响研究的结论。当被试取样分别为 10 个、20 个、40 个儿童时，三种取样情况下都得到显著的字的类型的主效应。但多重比较结果显示，取样大小会影响某些字的类型之间的关系效应。对于类型 1 与类型 3、类

型 2 与类型 3 之间的多重比较，三种取样情况下的结果是稳定的：儿童对第一类汉字的学习和记忆显著高于对第三类汉字的学习和记忆。同样，儿童对第二类汉字的学习和记忆显著高于对第三类汉字的学习和记忆。然而，取样大小影响了类型 1 与类型 2 的比较结果。当只取 10 个被试时，儿童学习和记忆第一类汉字的正确率显著低于第二类汉字。而当被试取样增加到 20 个和 40 个时，儿童学习和记忆第一类与第二类汉字之间的差异消失。表明在少数取样情况下得到的效应并不稳定可靠。

表 13-22　三种取样情况下多重比较的结果

被试数/人	平均数（经 arcsin 校正）	F	p	Newman-Keuls 检验
10	类型 1=1.169	50.714	0.000 0	类型 1<类型 2
	类型 2=1.446			类型 1>类型 3
	类型 3=0.603			类型 2>类型 3
20	类型 1=1.266	120.526	0.000 0	类型 1>类型 3
	类型 2=1.361			类型 2>类型 3
	类型 3=0.507			
40	类型 1=1.313	232.429	0.000 0	类型 1>类型 3
	类型 2=1.311			类型 2>类型 3
	类型 3=0.431			

在儿童反应时实验中，由于条件限制，通常可使用的被试数会少于一般的行为实验（如纸笔测验）。但由于儿童反应的变异较大，相对于成人反应时实验，儿童反应时实验仍然需要稍多一些的被试。例如，在快速命名实验中考察儿童形声字读音的规则性、一致性效应及其发展，研究者对三个年级的儿童命名规则—一致字、规则—不一致字、不规则—不一致字的命名反应时进行分析。其中，规则—一致字（类型 1）与规则—不一致字（类型 2）的反应时差异代表一致性效应，规则—不一致字（类型 2）与不规则—不一致字（类型 3）的差异代表规则性效应。

3（年级）×3（字的类型）的方差分析显示，字的类型的主效应显著，且与年级存在交互作用。不同取样情况下进行的多重比较结果见表 13-23。当被试取样较少时，如各年级被试分别为 10 个或 15 个儿童时，

尽管平均数显示出一定的效应趋势，但三年级的规则性效应、一致性效应以及四年级的一致性效应都未能达到显著水平。而在各年级被试增至 20 人的情况下，这两种效应才稳定出现。

表 13-23 三种取样情况下的多重比较结果

各年级被试数/人	年级	平均数/ms	F	p	Newman-Keuls 检验
10	三	类型 1=590	1.76	0.182	
		类型 2=586			
		类型 3=616			
	四	类型 1=577	5.60	0.006	类型 2<类型 3（规则性效应）
		类型 2=596			
		类型 3=635			
	六	类型 1=639	27.62	0.000	类型 1<类型 2
		类型 2=673			（一致性效应，p=0.10）
		类型 3=764			类型 2<类型 3（规则性效应）
15	三	类型 1=665	2.42	0.098	
		类型 2=640			
		类型 3=691			
	四	类型 1=650	3.81	0.026	类型 2<类型 3（规则性效应）
		类型 2=676			
		类型 3=712			
	六	类型 1=641	17.27	0.000	类型 1<类型 2（一致性效应）
		类型 2=686			类型 2<类型 3（规则性效应）
		类型 3=771			
20	三	类型 1=688	8.70	0.000	类型 2<类型 3（规则性效应）
		类型 2=667			
		类型 3=750			
	四	类型 1=652	7.22	0.001	类型 1<类型 2（一致性效应）
		类型 2=686			类型 2<类型 3（规则性效应）
		类型 3=731			
	六	类型 1=665	19.85	0.000	类型 1<类型 2（一致性效应）
		类型 2=708			类型 2<类型 3（规则性效应）
		类型 3=793			

在成人反应时实验中，通常误差变异较小，使用较少的被试就可以发现稳定的处理效应。当实验是项目内设计的（详见第八章），刺激材料的版本（version）较多时，考虑到整体实验规模，可适当减少各版本测试的被试量，每个版本取 10～12 个被试往往就能够得到比较可靠的效应。当实验中版本很少（如只有两个）时，可适当增加取样数量。例如，使用重复测量的实验设计，在启动命名实验中考察同音词及语义相关效应，研究者对不同启动条件下的目标词命名反应时进行分析。实验中的目标词在各种处理条件下是相同的，启动词 1 为目标词的同音词，启动词 2 为目标词的语义相关词，启动词 3 为控制条件（与目标词没有任何联系的词）。前两种启动词与控制条件的反应时差异代表同音启动及语义启动相应。

采用拉丁方设计将实验材料分为三组，形成三个版本。如果每个版本取 10 个被试，则整个实验有 30 个人的数据参与被试分析；若每个版本取 20 个被试，则整个实验有 60 个人的数据参与被试分析。如表 13-24 所示，两种取样条件下得到的实验效应完全相同，表明较少的被试（每个版本 10 人）能够反映出稳定的处理效应。

表 13-24 两种取样情况下的多重比较结果

被试数/人	平均数/ms	F	p	Newman-Keuls 检验
10×3	启动词 1=665	8.442	0.000 6	启动词 1<启动词 3
	启动词 2=676			启动词 2<启动词 3
	启动词 3=718			
20×3	启动词 1=635	15.456	0.000 0	启动词 1<启动词 3
	启动词 2=646			启动词 2<启动词 3
	启动词 3=675			

成人句子加工实验或眼动实验虽然也同样是反应时实验，但由于这类实验的处理效应较小或误差变异较大，通常需要较多的被试。例如，使用重复测量的实验设计，在移动窗口阅读条件下考察中文阅读中的字形及句法违反效应，研究者对不同违反条件下的关键词阅读反应时进行分析。关键词 1 为句子中的正确单字动词，关键词 2 为同音动词，关键词 3 为同音、不同词类的词（非动词）。关键词 2、3 分别与正确动词的阅读反应时

差异代表两种违反效应（字形违反，及“字形＋词类”违反）。

采用拉丁方设计将实验材料分为三组，形成三个版本。如果每个版本取 10 个被试，则整个实验有 30 个人的数据参与被试分析。如果将每个版本的被试增加到 16 个人，则整个实验有 48 个人的数据参与被试分析。表 13-25 显示，当每个版本只有 10 个人时，关键词类型效应不显著，两种违反都没有引起阅读反应时的显著增加。而当各版本的人数增加至 16 人时，关键词的类型效应达到边缘显著，且“词类＋字形”违反效应也边缘显著。因此在这类实验中，保证足够的被试取样对于实验结果至关重要。在很多前人的研究中，每个版本需要 20 人或更多的被试才能得到比较稳定的处理效应。

表 13-25　两种取样情况下的多重比较结果

被试数/人	平均数/ms	F	p	Newman-Keuls 检验
10×3	关键词 1＝269	0.297	0.745 9	
	关键词 2＝275			
	关键词 3＝277			
16×3	关键词 1＝268	2.639	0.076 2	关键词 1＜关键词 3
	关键词 2＝281			（$p<0.1$）
	关键词 3＝294			

进行检验力分析是任何一个实验计划中必须的一个步骤，估计检验力就是我们在某种程度上控制Ⅱ型错误。这就是说，我们需要在实验实施前进行一个关于处理效应，以及误差大小的合理估价。然而，在很多情况下，尤其是对我们不熟悉的研究，对这些信息作事先的、可靠的估计几乎是不可能的。研究者常用的一种可行的方法是做一个微型实验或者预实验以提供这样的估计。靠预实验中提供的信息我们可以确定我们的实验所需要的检验力，确定实验所需要的被试数量。有时我们会发现增加被试的代价太大，是很难实行的。我们还可以考虑通过其他方法增加检验力。例如，在实验实施中加大控制，选择能使处理效应达到最大化的自变量水平，选择一个更敏感的实验设计，或者选用一个能进行协方差分析的协变量等。统计学家已经提供了多种设计和分析数据的方式，但要看我们怎样

有效地使用它们。

（四）检验力和独立重复

假定我们将同样的实验重复若干次，每次都不能拒绝虚无假说，然而，每次都重复得到同样的结果模式。在这种情况下，我们应当如何看待这样的结果模式？我们知道，F 检验是不考虑这些独立实验中数据的一致性的。那么，有多大的可能性我们在偶然概率上获得同样的结果模式？答案是：几乎是不可能的。其实当这些实验是独立进行的，并独立重复了实验结果，即使统计结果不显著，其结果也可能向研究者提供了非常重要的信息，甚至可以说比某些实验一次获得的 F 显著性提供了更多的东西。

为什么独立重复实验是需要的？我们将区分两种重复实验，一种是计划作为实验一部分的独立重复，另一种是没有计划的。人们将重复实施若干次实验纳入自己的研究计划，其中一个可能的原因是，他一次可能只能施测少量的被试。例如，在动物学习实验中，每个动物要进行过度训练，持续几周。另外，在教学方法的研究中，可能教学要持续一个学期，实验持续时间长可能使研究者一次不能施测大量被试。再如，有限的空间使一次施测的被试有限。在上述情况下，研究者在设计实验时便将重复实施若干次实验纳入其研究计划。重复实验的另一个原因是研究者可能想通过重复实验考察多个无关因素的作用，如实验者的施测、测验房间、动物的批次、测验时间、不同的学校和班级等对实验结果的影响。因为上述任何一个因素在同一实验内可能是恒定的，但在重复实验时都可能改变。在重复实验时改变这样的因素，使实验之间无关因素的差异尽可能大，对研究者是一件好事。这种方法的好处是移去了变异的误差源，这些误差源在没有重复实验时可能是难以控制的。对更多的研究者来说，重复实验不是预先计划好的，而是研究者希望通过重复或部分重复前人的实验，观察处理效应是否在不同的研究里可以得到重复，以检验前人研究结论的正确性。

（五）增加检验力

以上可以看出，检验力分析的主要目的是确定要达到预定的可能存在的处理效应值所需的被试数量，或者说确定在预定的 α 水平上没有发现处理效应是否可能主要由于被试数量不够。检验力的水平依赖于几个因素：（1）用来确定显著性水平的统计方法；（2）选择的 α 水平；（3）实验中使

用的样本大小；(4) 要研究的处理效应量的大小。因此，增加检验力一个方法是增加处理效应量，另一个方法是减小误差变异。

下面介绍一种确定处理效应量的大小的方法［eta(η) 的计算］和一种增加检验力的方法（区组方法）。

例如，研究者要探讨一种新的治疗方法的疗效。在一个完全随机实验设计中，研究者首先选取了 10 个病人，并随机分配 10 个病人接受两种治疗方法：新的治疗方法和传统治疗方法，接受每种治疗方法的病人各半。研究者还可以使用随机区组实验设计。在这种设计中，研究者首先将 10 个病人按其焦虑程度分为五组，每组中的两个病人随机分配接受两种治疗之一。两种实验设计及数据结果与分析如下：

表 13-26　完全随机设计

	治疗	控制
	10	12
	6	8
	4	7
	2	3
	1	2
$\sum$	23	32
$\overline{X}$	4.6	6.4

表 13-27　随机区组设计

焦虑程度	治疗	控制	$\sum$
很高	10	12	22
高	6	8	14
中	4	7	11
低	2	3	5
很低	1	2	3
$\sum$	23	32	
$\overline{X}$	4.6	6.4	

表 13-28　完全随机方差分析表

变异来源	SS	df	MS	F	p	eta
组间	8.1	1	8.1	0.56	0.48	0.26
组内	116.4	8	14.6			

表 13-29　随机区组方差分析表

变异来源	SS	df	MS	F	p	eta
处理	8.1	1	8.1	23.14	0.009	0.92
区组	115.0	4	28.6	82.14	0.000	0.99
残差	1.4	4	0.35			

比较两种实验设计的数据计算，可以看到，即使在原始数据完全相同的情况下，两种方差分析的结果也是很不相同的。使用完全随机方差分析表明两种治疗方法的差异是不显著的，$F(1, 8)=0.56$，$p=0.48$。然而使用随机区组方差分析表明，两种治疗方法是差异显著的，$F(1, 4)=23.14$，$p=0.009$。为什么两种方差分析的结果会出现如此大的差异？从表 13-28 和表 13-29 中可以看到，表中的组间或处理均方是相同的，然而误差变异非常不同，完全随机方差分析表中的误差均方是 14.6，而随机区组方差分析表中误差均方是 0.35。

我们再来看一下两个实验的处理效应量的大小。eta 是衡量处理效应量大小的一个指标（Rosenthal & Rosnow，1991)，它可以用以下公式计算：

$$\text{eta}=\sqrt{\frac{SS_{组间}}{SS_{组间}+SS_{组内}}}$$

在完全随机实验中，处理效应量大小的计算是：

$$\text{eta}=\sqrt{\frac{8.10}{8.10+116.40}}=0.26$$

在随机区组实验中，处理效应量大小的计算是：

$$\text{eta}=\sqrt{\frac{8.10}{8.10+1.40}}$$

$$=0.92$$

区组效应的计算：

$$\text{eta}=\sqrt{\frac{115.0}{115.0+1.40}}$$

$$=0.99$$

如果研究者希望增加完全随机实验设计的有效性，可以通过增加被试数量的方法。以上面的实验为例，已知随机区组实验的处理效应量大小为 0.92。如果我们希望使用完全随机实验达到相同的处理效应量，所需要的被试数量计算如下：

$$\text{reps}=\frac{MS_{组内}\times 区组数量}{MS_{残差}}=\frac{14.6\times 5}{0.35}=208.6$$

计算表明要在完全随机实验设计中增加统计检验的显著性，如果保持处理效应量的大小不变，需要在完全随机实验设计中增加到 209 个被试，以达到随机区组实验设计中 10 个被试得到的统计显著性。

另一种方法，我们也可以通过增加处理效应的方式，增加完全随机设计的统计检验显著性。如果保持被试数量不变，需要将完全随机实验设计中的处理效应增加到 337.9，以达到随机区组实验设计中处理效应为 8.1 时得到的显著性，则增加处理效应的计算为：

$$SS=208.6\times(4.6-5.5)^2+208.6\times(6.4-5.5)^2=337.9$$

表 13-30 完全随机方差分析表

	SS	df	MS	F
组间	337.93	1	337.93	23.15
组内	116.40	8	14.60	

本章主要观点

• 转换是指对一组数据进行系统的转变。在行为科学研究中，需要对原始数据进行转换的原因可能包括：原始数据不能很好地满足 F 检验的需要，原始数据的变异大，因而影响对处理效应的估计。常用的数据转换方法包括平方根转换、log 转换、角度转换等。

• 选择合适的转换方法，要考虑原始数据的性质、分布及数值的变异

情况等。

• 可以对原始数据进行数据转换的原理是，如果数据呈正偏态分布或者数据变异较大导致方差分析不显著，是与人为选择因变量测量量表的类型有关的话，那么数据的变异较大、正偏态分布等状况是可以通过数据转换或改变测量量表得到改善的。

• 处理不等组数据的两种统计方法是无权重的分析和权重的分析。在无权重的分析中，每个处理平均数对确定平方和 SS 有同等的贡献。而在有权重的分析中，每个处理平均数对确定平方和 SS 的贡献与每个组被试的数量有关。使用无权重或权重分析的前提是假定不等组的出现是与实验处理条件分配无关的。

• 实验的敏感性的量化指标是检验力。检验力指当备择假说为真时，拒绝虚无假说的可能性。或者说，检验力是当虚无假说为假，作出正确决策的可能性。

• 增加实验的检验力通常使用的方法包括：增加每种实验条件下观察的数量，分离出未控制的变量或无关变异源，增大处理效应。

• 在计划实验和解释统计不显著的实验结果的时候，对检验力的理解是非常重要的，它可以帮助研究者进一步分析结果不显著的原因和如何改进实验设计。

思考题

1. 如何处理极端数据、缺失数据等问题？

2. 什么是数据转换？为什么要进行数据转换？有哪些常用的数据转换方法？

3. 如何选择合适的数据转换方法？

4. 不同处理条件下被试的数量不相等，对推论统计可能有哪些影响？

5. 什么是处理不等组数据的无权重的分析？什么是权重的分析？

6. 什么是检验力？有哪些途径可以提高检验力？

第十四章
方差分析与多重回归模型

前面的章节中，我们主要介绍了各种实验设计及其与方差分析统计方法的结合。方差分析是在实验设计课程中最常介绍的统计方法。在使用这种统计方法时，对实验设计中的自变量有一些比较严格的要求。

第一，自变量的水平应当是分类的，而且水平的数量是有限的。即使当一些自变量的水平本身是数量化的、连续的，也需要将它们转换成分类的。例如，要探讨字词频率对汉字命名反应时的影响，汉字频率是一个连续变量，可能从每百万次语料中出现 0 次到每百万语料中出现数千次。对这样的自变量，需要首先将其变为一个分类变量，如将汉字频率的变量定义为高频和低频，将若干个不同频率范围的字归入两个类别，如高频字的范围为每百万字 500～1 000 次，低频字的范围为每百万字 1～10 次。一般情况下，我们不能直接使用每一个汉字的频率作为自变量的水平，以探讨汉字频率变化对命名反应时的影响。

第二，在使用方差分析的实验设计中，自变量与自变量之间是互相独立的，一般是通过实验设计使两个因素的处理水平相互正交。例如，研究者要探讨记忆编码方式与提取方式对记忆成绩的影响，编码方式分为简单重复编码（A1）、语音编码（A2）和语义编码（A3）三种方式，提取方式分为再认（B1）和回忆（B2）两种方式。这是一个 2×3 的实验设计，处理水平的结合有 A1B1、A1B2、A2B1、A2B2、A3B1 和 A3B2。其中六个单元之间是相互独立的，一般情况下，研究者会在每个单元中使用同样数量的实验项目。

第三，在使用方差分析的实验设计中，自变量及其水平都是有限的。一方面，如果自变量的数量多于五个，多次交互作用是很难解释的。例如，研究者要探讨图片命名一致性、可表象性、熟悉性、线条复杂性、口

语获得年龄对图片命名反应时的影响，如果使用方差分析方法，获得的五次交互作用往往是非常难以解释的。如果研究者得到一个显著的五次交互作用，可能需要还原成简单简单简单简单效应，才能清楚地解释这个五次交互作用的含义。另一方面，如果自变量的水平数量很多，也会带来结果难以解释的问题。例如，要探讨字词频率对汉字命名反应时的影响，如果直接使用每一个汉字的频率作为自变量的水平，实验设计和结果解释都是难以完成的。

方差分析对实验设计中的变量有比较严格的要求，而很多心理、教育、社会实验很难满足这样的要求，因而在一定程度上限制了它的使用。多重回归是一种比方差分析更加一般的数据处理途径，它可以与更加灵活的实验设计相结合，探讨多个自变量与一个因变量之间的关系。在固定效应模型中，任何固定效应的方差分析都可以写成回归模型分析的形式。然而，回归模型不一定可以用方差分析代替。

本章介绍回归模型与实验设计的关系，从一般线性模型的角度看回归模型与方差分析模型的异同。

第一节　实验设计与多重回归模型分析

一、与多重回归模型分析相结合的实验设计的特点

近年来，在发展心理、教育心理、社会心理等许多领域，研究者越来越多地在实验研究中使用多重回归模型（multiple regression model）分析数据，它的优点如下。

（1）多重回归模型中自变量之间可以存在相关，而且不需要通过实验设计的方法将两个因素完全独立。例如，要研究儿童智力与口语能力对阅读理解成绩的影响，智力（A）与口语能力（B）可能本身是存在相关的。在方差分析中，研究者需要通过实验设计来获得两个独立的因素。例如，研究者利用智力测验和口语词汇测验选取了四组儿童：（1）智力高、口语能力高；（2）智力高、口语能力低；（3）智力低、口语能力高；（4）智力低、口语能力低。每组各 50 名儿童参加实验。

表 14-1 智力和口语能力不同的四组儿童

	智力高（A1）	智力低（A2）
口语能力高（B1）	50	50
口语能力低（B2）	50	50

从表 14-1 中可以看出，通过设计四个处理水平的结合，得到的两个因素对阅读理解的影响是完全独立的。然而在实际中，我们不容易得到智力低、口语能力高的被试。即使这四种类型的儿童都可以找到，他们在现实中的分布可能也很不均匀。在多重回归模型中，可以比较好地解决这样的问题。研究者可以按照自然的情况对被试进行编码，不需要两个因素完全独立，被试在各种条件下也不需要是等组的。甚至某种条件下的被试可能是缺失的。

（2）多重回归模型中自变量的水平可以是分类的，也可以是连续数量的，水平的数量也是没有严格限定的。在一个回归模型中有七八个自变量，甚至十几个自变量是非常普遍的。自变量如果是智力，可以直接使用每个被试的智力分数作为智力水平的编码，而不需要区分为高智力、低智力或其他的分类。多重回归模型分析更加关心的是各种因素的主效应的相对影响，或者是某一个因素是否独立对因变量产生影响。虽然多重回归模型中也可以计算交互作用，但是对交互作用的解释，特别是对多次交互作用，或对含连续自变量的交互作用的解释也是比较困难的。

（3）多重回归模型尤其可以用于无法进行严格实验设计的情况下。在方差分析中，一般要求各处理条件下的被试数、项目数是相等的。然而在多重回归模型中，实验处理结合的单元可以是不平衡的。

我们举一个例子来说明多重回归模型的优越性。例如，研究者要探讨在文章阅读过程中的字词识别特点。研究中使用移动窗口技术呈现文章，每次呈现一个词。被试的任务是自定步速阅读。计算机记录文章中每一个词的平均阅读时间（毫秒）作为因变量。阅读的文章是八篇科学短文。在文章阅读过程中，影响词的阅读时间的因素可能有许多，研究者选取的自变量有六个因素：字的笔画数（词中字的平均笔画数）、词频、词的长度（词中的字数）、词的概念重要性、句子界限、文章难度（舒华等，1996）。

可以看出，要研究比较自然状态下的文章阅读加工过程，是几乎不可能使用方差分析统计方法的。第一，当科学短文选定后，文章中所包含的字词、句子的特性随之而定。研究者要分析的是被试对文章中所有词的阅读时间，一旦文章确定了，就不可能对字词作进一步严格的选择、设计了。第二，六个自变量中有些自变量水平是分类的，如词的长度可以分为单字词、双字词、多字词。词的概念重要性可以分为科学概念名词、一般概念词。句子界限可以分为词在句子中间、在逗号尾、在句号尾或在段落尾。文章难度可以分为科普文章、专业技术文章。然而，还有些自变量水平是连续的，如字的笔画数可以在 1～15 笔画之间变化。词频可以在每百万 0～1 000 次之间变化。第三，如果记录和分析真实文章中所有的字词，而不是个别目标词的阅读时间，明显不可能做一个整齐的 3（单字词、双字词、多字词）×2（科学概念名词、一般概念词）×4（句子中间、逗号尾、句号尾、段落尾）×2（科普文章、专业技术文章）×2（字的平均笔画数多、字的平均笔画数少）×2（高频词、低频词）的实验设计。或者说如果选用真实的短文，不可能保证各个处理结合单元中的词数是平衡的。

使用多重回归模型进行分析，研究者按照自然的情况对被试进行了编码，即对六个自变量中词的长度、词的概念重要性、句子界限、文章难度使用分类编码，对字的笔画数、词频使用连续编码，数据分析得出了很有意义的结果。结果表明，在字词水平的特性上，字的平均笔画数的效应是不显著的，词频（$p<0.01$）和词的长度（$p<0.01$）的效应是显著的。描述统计的平均数分析表明，较低频词（680 毫秒）比较高频词（539 毫秒）阅读时间长。多字词（644 毫秒）比双字词（564 毫秒）和单字词（521 毫秒）阅读时间长。在语篇水平，词的概念重要性（$p<0.01$）、句子界限（$p<0.01$）、文章难度（$p<0.01$）的效应是显著的。平均数分析显示，科学概念名词（658 毫秒）比一般概念词（547 毫秒）的阅读时间长。与句子中间词（545 毫秒）相比，在逗号尾（673 毫秒）、句号尾（724 毫秒）或段落尾（881 毫秒）的词的阅读时间要长得多，表明在逗号尾、句号尾或段落尾的词的阅读时间还包含了进行句子整合或文章整合的时间。在科普文章中词的平均阅读时间是（514 毫秒），而在专业技术文

章中词的平均阅读时间是（626 毫秒），表明阅读过程中词的阅读时间受到文章水平特征的影响。回归分析还发现了一些显著的交互作用。句子界限和词频的交互作用（$p<0.01$）是显著的。进一步的平均数分析表明，在句子中间时，较低频词（643 毫秒）比较高频词（520 毫秒）阅读时间长得多，而较低频词与较高频词阅读时间的差距在逗号尾［682－620＝62（毫秒）］和句号尾处［767－748＝19（毫秒）］大大减少。可以看出，在回归分析中，即使可以使用水平是连续的自变量（如词频），但是当用文字对主效应和交互作用进行解释时，还是倾向于用分类定义（较高频词、较低频词）来叙述。

二、方差分析与多重回归模型的关系

以前一般认为，回归分析（regression analysis）是在相关分析的基础上建立数学表达式，来确定变量之间关系的模型。研究结果可以得到变量之间的相关，但不能作出因果解释。以前大多数关于回归分析和方差分析的讨论更多地强调它们在探讨自变量和因变量关系上存在的差异。在本章的介绍中，我们将可以看到，方差分析是回归分析的一种特例，与实验设计相结合的回归分析也可以在一定程度上探讨因果关系。重要的问题在于实验设计。实验研究中探讨变量间因果关系的核心问题是操纵、改变自变量，并观察其对因变量的影响。因此，与方差分析类似，如果实验中没有操纵、改变自变量，或自变量均为不可操纵的因素，如年龄、性别等，回归分析探讨的是相关关系。而如果在研究中包含了自变量的操纵，回归模型分析数据也是可以探讨因果关系的。

在一个方差分析研究情景下，实验者对在何种程度上一个或多个定量的或定性的自变量的变异可以解释定量的因变量的变异感兴趣。在方差分析的实验研究中，自变量是事先确定的，每个自变量水平或者是被操纵的，或者是事先选择的。然而，在方差分析中，虽然自变量可以是定量的，但实际上，由于方差分析特点的限定，实验设计时定量的自变量会被转化为分类的变量，因此不能作出有关定量自变量与因变量之间关系的实质的假设。或者说，方差分析方法忽略了自变量水平之间量的差异的大小，以及自变量水平的量与因变量变异之间关系的实质。与方差分析相

比，回归模型分析在探讨定量自变量水平与因变量变异之间的关系上有一定的优势。

回归模型分析有预测和解释两种功能：预测功能主要关心自变量对因变量的预测程度；解释功能主要关心由实验处理引起的因变量变化的百分数与误差引起的因变量变化相比是否差异显著，其思想与在方差分析中平方和分解的思想是一致的。后者是本书关心的重点。

第二节　回归分析的预测功能

在一个回归分析的研究中，实验者对一个或多个定量的自变量与定量的因变量之间关系的大小感兴趣，换句话说，实验者想知道自变量的变异在何种程度上可以预测或解释因变量的变化。在回归分析的预测和解释两种功能中，第一种功能是大家比较熟悉的，即当研究者对于自变量对因变量的预测程度感兴趣时，研究者更关心所有自变量的变异在多大程度上可以预测因变量的变化。在这一节中，我们举例简要介绍回归分析的预测功能。将回归分析用于预测时，研究者期望找到反映自变量与因变量之间关系规律性的数学表达式，基本方法是根据一组数据，确定自变量与因变量之间定量的直线关系，然后对这个关系进行可靠程度的检验，所得到的可靠的关系式可以用来进行预测（程书肖和李仲来，1988；冯伯麟，2005）。

我们知道，在一元线性回归中，回归方程的公式是：

$$\hat{Y}=a+bX$$

其中 X 和 Y 分别代表自变量与因变量；$\hat{Y}$ 是对应于 X 的因变量的估计值；a 是直线在 Y 轴上的截距，或者说当 X 等于 0 时的 Y 值；b 是直线的斜率，或回归系数，表示当 X 变化一个单位，Y 将变化 b 个单位。

如果能通过收集的一组数据计算出 a 和 b 的值，回归方程就建立了。我们介绍一个用于预测的回归方程的研究例子。文章的易懂性评定是心理学家和教育学家共同关注的问题。文章的易懂性实际上是指文章阅读的难度，它通常是一个比较主观的指标。如果能够建立一个易懂性评定公式，客观地评价各种文章的难度，可以为学校课本中课文选定、课外读物的编

写、语文考试中阅读文章的选择等提供客观的检测指标和快捷的评定方法。在一个汉语文章易懂性的研究中，研究者关心特定语文水平的汉语读者在阅读文章时，影响读者成功地阅读理解内容的各种因素（孙汉银，1985）。易懂性公式就是从文章中选出一些对阅读理解有重要影响并且能够量化的因素作为自变量，以读者的理解分数作为因变量而建立的回归方程。利用回归方程可以便捷地计算出一篇文章的难度水平。

研究者分析了影响汉语易懂性的因素。他首先在报刊、杂志、课本和书籍等材料中随机挑选了 20 篇字数在 250 个字左右的段落，其中说明文 2 篇，政论性文体 5 篇，科普读物 2 篇，散文 5 篇，记叙文 6 篇。被试为初中二年级学生，共 150 人。将 20 篇材料随机排列，用完形填空（cloze procedure）的形式，每隔五个词删除一个词，并要求被试进行完形填空。被试写出的完全合乎标准答案的词占总的被删除词的比率作为正确率。用正确率的百分位数作为因变量，或文章的理解分数。实验中共有 13 个可能影响阅读理解分数的因素，将其作为自变量：X_1＝总笔画数；X_2＝多笔画字数；X_3＝难字总数；X_4＝难词总数；X_5＝以逗号为起点的句子总数；X_6＝以句号为起点的句子总数；X_7＝字均笔画数；X_8＝难字百分比；X_9＝难词百分比；X_{10}＝句均字数（逗号）；X_{11}＝句均词数（逗号）；X_{12}＝句均字数（句号）；X_{13}＝句均词数（句号）。然后利用多元回归中的逐步分析后退法（backward）或逐步递减法计算最佳的回归方程。

第一步，研究者让全部 13 个自变量都进入易懂性公式，得到的公式是：

$$\begin{aligned}Y=&71.55496+0.01789X_1-0.90679X_2+0.21234X_3-\\&0.000351779X_4-2.07337X_5-1.06465X_6+\\&23.01332X_7-1.2832X_8-2.72230X_9-7.51913X_{10}-\\&0.35235X_{11}-4.75818X_{12}+6.26253X_{13}\end{aligned}$$

计算结果表明，该公式能够解释文章理解分数全部变异的 80.768%，但方程不显著。

第二步，研究者删除了 X_4（难词总数），得到的公式是：

$$\begin{aligned}Y=&71.54905+0.01789X_1-0.90677X_2+0.2123X_3-\\&3.07327X_5-1.06466X_6+23.01539X_7-1.28313X_8-\\&2.72297X_9-7.51759X_{10}-0.35405X_{11}-4.75815X_{12}+\end{aligned}$$

$6.262\ 47X_{13}$

计算结果表明，该公式可以解释全部变异的 80.768%，但方程不显著。

第三步，研究者继续删除 X_{11}（以逗号为起点的句均词数），方程仍然不显著。

第四步，研究者删除 X_3（难字总数）时，该公式能解释全部变异的 80.758%，方程显著。

第五步，删除 X_6（以句号为起点的句子总数），该公式能解释全部变异的 80.478%，方程显著。

第六步，删除 X_1（总笔画数），该公式能解释全部变异的 79.72%，方程显著。

第七步，删除 X_8（难字比例），该公式能解释全部变异的 78.845%，方程显著。

第八步，删除 X_2（多笔画字总数），该公式能解释全部变异的 76.487%，方程显著。

第九步，删除 X_5（以逗号为起点的句子总数），该公式能解释全部变异的 71.107%，方程显著。

第十步，删除 X_{10}（以逗号为起点的句均子数），得到的公式是：

$$Y=-7.006\ 85+14.345\ 87X_7-2.137\ 91X_9-3.387\ 99X_{12}+4.003\ 71X_{13}$$

该公式能解释全部变异的 71.080%，方程显著。

第十一步，由于 X_{12}与 X_{13}之间相关度很高（$r=0.968\ 8$），若强制性删除 X_{12}之后，该公式只能解释全部变异的 50.843%，方程显著。

第十二步，若保留 X_{12}，强制性删除 X_{13}，该公式能解释全部变异的 57.691%，方程显著。

可以看出，数据分析所得的方程中，效果最好的是第十个方程，其中公式中保留的四个因素是字均笔画数、难词百分比、句均字数（句号）、句均词数（句号）。这四个因素的结合可以预期阅读理解分数中 71%的变异，或者说汉语易懂性公式中字均笔画数（X_7）、难词百分比（X_9）、句均字数（句号）（X_{12}）和句均词数（句号）（X_{13}）四个因素的结合可以大体评价一篇汉语文章的难度。文章中难词的百分比、以句号为单位的句子中的字数、词数对阅读文章难度有重要影响，表明词是现代汉语中阅读理

解的基本单元。同时人们在阅读时需要对汉字进行识别，所以字均笔画数对理解分数的影响力仍然是很大的。

易懂性公式预测的准确性如何，或者说公式的预测效度如何？实验结束之后，研究者选择了另外 20 篇文章（其中说明文 2 篇，政论性文体 2 篇，科普读物 2 篇，小说 2 篇，散文 4 篇，记叙文 8 篇），请中小学教育专家、中小学语文教师和编辑人员进行文章适读水平的评价，所得到的适读水平与易懂性公式中吻合度最高的是：

$$Y=16.30274+10.2539X_7-1.67258X_9-0.71427X_{12}$$

可以看到，专家评价得出的公式与易懂性公式非常接近，两个公式中均包含了字均笔画数（X_7）和难词百分比（X_9）两个因素。稍有不同的是，易懂性公式中句均字数（句号）和句均词数（句号）均包括在内，而专家评价的公式中只保留了句均字数（句号）（X_{12}）。

通过上面的例子可以看到，利用回归方程可以便捷地计算出一篇文章的难度水平。在这样的研究中，研究者对探讨对因变量预测程度最好的自变量及其组合感兴趣。

第三节　多重回归模型与方差分析实验设计模型

回归方程的第二种功能是本章中要重点介绍的，即用回归分析解释自变量与因变量的关系时，研究者更关心不同的自变量是如何影响因变量的，每个自变量或自变量之间的交互作用可以解释因变量变化的程度如何。当使用回归分析的解释功能时，回归分析和方差分析是非常接近的。当回归分析研究中的自变量或部分自变量的水平是被研究者操纵改变的，或者是事先选择的，回归分析研究同样可以探讨自变量与因变量之间的因果关系。所以，研究要得到因果结论，重要的不是使用方差分析或回归分析处理数据，而是在于实验设计中是否包含可操纵改变的自变量，或者是否可以使研究者在一组固定的 X 值的结合上观察因变量 Y 值。回归分析中自变量可以是定量（quantitative）的，也可以是定性（qualitative）的，如性别、治疗的种类；定量的自变量可以是连续的，也可以是分类的。在很多研究情况下，回归分析可以与更加灵活的实验设计相结合。在一些实

验情境下，回归分析是唯一的选择。

这一节中我们将简要介绍回归模型的解释功能，即用回归分析探讨自变量的主效应或自变量之间的交互作用可以解释因变量变化的程度。我们通过将回归模型和方差分析实验设计模型的矩阵表达相比较，了解矩阵表示在两种模型中是相同的。这种比较会使我们更清楚地理解为什么说当关注回归分析的解释功能时，回归分析和方差分析是非常接近的（Kirk，1982）。

在一元线性回归中，回归方程的公式是：

$$\hat{Y}=a+bX$$

当研究中有多个自变量时，我们对多个自变量 X_1，X_2，X_3，…，X_k 共同对因变量 Y 的线性影响感兴趣，这叫做多重回归分析。在一个研究中，通常需要进行 n 次观测，得到 n 组观测数据。假设数据如下：

$$X_{11}, X_{12}, \cdots, X_1, Y_1$$

$$X_{21}, X_{22}, \cdots, X_2, Y_2$$

……

$$X_{n1}, X_{n2}, \cdots, X_{nk}, Y_n$$

多重回归可以从 n 组观测数据出发，建立因变量 Y 与自变量 X_1，X_2，X_3，…之间的线性关系。

多重回归模型的公式是：

$$Y_i=\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\varepsilon_i \ (i=1, 2, 3, \cdots, n)$$

公式中 Y_i 是第 i 个实验单元（或被试）的因变量观测值，n 表示 n 组观测数据或被试数。可以看到，Y 的观察值由两部分组成：恒定的预期项 $\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots$ 和随机误差项 ε_i。其中误差项 ε_i 反映了因变量观测值 Y 中不能被自变量解释的部分。由于误差项是随机变量，因此 Y 也是随机变量。模型中 X_{i1}，X_{i2}，…是自变量的值，β_1，β_2，…是模型中自变量 X_i 的偏回归系数（partial regression coefficient）。一个偏回归系数指当其他自变量不变时，它所说明的自变量变异引起的因变量的变异。ε_i 是误差项，它是一个平均数为 0，变异为 σ^2 的随机变量。当模型中只有一个自变量时，是简单回归模型（simple regression model）。当模型中有两个或多个自变量时，是多重回归模型（multiple regression model）。注意，X 是矩阵的一系列指示变量值，这种矩阵叫做结构矩阵（structural matrix）。

如果我们将公式 $Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\varepsilon_i$ 展开，代入（$i=1$，2，3，…，n），公式可以写做：

$$Y_1=\beta_0+\beta_1X_{11}+\beta_2X_{12}+\cdots+\beta_kX_{1k}+\varepsilon_1$$

$$Y_2=\beta_0+\beta_1X_{21}+\beta_2X_{22}+\cdots+\beta_kX_{2k}+\varepsilon_2$$

……

$$Y_n=\beta_0+\beta_1X_{n1}+\beta_2X_{n2}+\cdots+\beta_kX_{nk}+\varepsilon_n$$

公式也可以以矩阵的方式表示：

$$\begin{bmatrix}Y_1\\Y_2\\Y_3\\\vdots\\Y_n\end{bmatrix}=\begin{bmatrix}1&X_{11}&X_{12}&X_{13}&\cdots&X_{1k}\\1&X_{21}&X_{22}&X_{23}&\cdots&X_{2k}\\1&X_{31}&X_{32}&X_{33}&\cdots&X_{3k}\\\vdots&\vdots&\vdots&\vdots&&\vdots\\1&X_{n1}&X_{n2}&X_{n3}&\cdots&X_{nk}\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\\\beta_2\\\vdots\\\beta_k\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\vdots\\\varepsilon_n\end{bmatrix}$$

公式还可以抽象地表示为：

$$Y=X\times\beta+\varepsilon$$

我们还记得第九章中介绍的单因素完全随机实验设计模型的写法：

$$Y_{ij}=\mu+\alpha_j+\varepsilon_{i(j)}$$

$$(i=1,\ 2,\ \cdots,\ n;\ j=1,\ 2,\ \cdots,\ p)$$

模型中的 Y_{ij} 表示实验中第 i 个被试在第 j 个处理水平上的观测值。μ 表示总体平均数，它是未知的，但可以用样本的总平均数$\overline{Y_{..}}$来估价。α_j 表示水平 j 的处理效应，它是通过样本 $Y_j-\overline{Y_{..}}$ 来估价的。$\varepsilon_{i(j)}$ 表示误差变异，是用 $Y_{ij}-\overline{Y_{.j}}$来估价的。

如果我们展开单因素完全随机实验设计模型：

$$Y_{11}=\mu+\alpha_1+\qquad\qquad+\varepsilon_{1(1)}$$

$$Y_{21}=\mu+\alpha_1+\qquad\qquad+\varepsilon_{2(1)}$$

$$Y_{12}=\mu+\qquad\alpha_2+\qquad+\varepsilon_{1(2)}$$

$$Y_{22}=\mu+\qquad\alpha_2+\qquad+\varepsilon_{2(2)}$$

……

$$Y_{1p}=\mu+\qquad\qquad\alpha_p+\varepsilon_{1(p)}$$

$$Y_{2p}=\mu+\qquad\qquad\alpha_p+\varepsilon_{2(p)}$$

其中，Y_{11} 和 Y_{21} 表示接受 α_1 水平的两个被试的观测值，Y_{12} 和 Y_{22} 表示接受 α_2 水平的两个被试的观测值。

我们还可以将公式用系数的方式表示：

$$Y_{11}=1\times\mu+1\times\alpha_1+0\times\alpha_2+\cdots+0\times\alpha_p+\varepsilon_{1(1)}$$
$$Y_{21}=1\times\mu+1\times\alpha_1+0\times\alpha_2+\cdots+0\times\alpha_p+\varepsilon_{2(1)}$$
$$Y_{12}=1\times\mu+0\times\alpha_1+1\times\alpha_2+\cdots+0\times\alpha_p+\varepsilon_{1(2)}$$
$$Y_{22}=1\times\mu+0\times\alpha_1+1\times\alpha_2+\cdots+0\times\alpha_p+\varepsilon_{2(2)}$$

……

$$Y_{1p}=1\times\mu+0\times\alpha_1+0\times\alpha_2+\cdots+1\times\alpha_p+\varepsilon_{1(p)}$$
$$Y_{2p}=1\times\mu+0\times\alpha_1+0\times\alpha_2+\cdots+1\times\alpha_p+\varepsilon_{2(p)}$$

进一步，可以将公式以矩阵的方式表示：

$$\begin{bmatrix}Y_{11}\\Y_{21}\\Y_{12}\\Y_{22}\\\vdots\\Y_{2p}\end{bmatrix}=\begin{bmatrix}1&1&0&\cdots&0\\1&1&0&\cdots&0\\1&0&1&\cdots&0\\1&0&1&\cdots&0\\\vdots&\vdots&\vdots&&\vdots\\0&0&0&\cdots&1\end{bmatrix}\begin{bmatrix}\mu\\\alpha_1\\\alpha_2\\\vdots\\\alpha_p\end{bmatrix}+\begin{bmatrix}\varepsilon_{1(1)}\\\varepsilon_{2(1)}\\\varepsilon_{1(2)}\\\vdots\\\varepsilon_{2(p)}\end{bmatrix}$$

最后可以抽象地表示为：

$$Y=X\times\alpha+\varepsilon$$

这样的矩阵也叫做结构矩阵。我们将多重回归模型的矩阵和完全随机实验设计模型的矩阵比较一下，可以看出，在线性模型的框架下，回归模型和实验设计模型是相同的。

当用回归分析解释自变量与因变量的关系时，与方差分析类似，研究者对每个自变量的主效应或自变量之间的交互作用的变异可以解释因变量变异的程度感兴趣。或者说，自变量主效应或交互作用引起的因变量变异与误差变异相比是否显著。这时，回归方程还可以写成以下形式：

$$\sum(Y_i-\overline{Y}_.)^2=\sum(Y_i-\hat{Y}_i)^2+\sum(\hat{Y}_i-\overline{Y}_.)^2$$

其中 $\sum(Y_i-\overline{Y}_.)^2$ 表示观测值 Y 的总的变化平方和，是总平方和 SS_T；$\sum(\hat{Y}_i-\overline{Y}_.)^2$ 表示回归方程估计值 $\hat{Y}_i$ 相对于总平均数 $\overline{Y}_.$ 的偏离，是回归

方程本身变化的平方和，又称做回归平方和 SS_R；$\sum(Y_i-\hat{Y}_i)^2$ 表示观测值 Y 相对于回归线的偏离，是可以预测的误差，又称做误差平方和 SS_E；

使用 F 检验时，总自由度是 $N-1$，回归自由度是 k，误差自由度是 $N-k-1$。其中 N 是被试或样本总数，k 是自变量的数量。

F 显著性检验的公式是：

$$F=\frac{SS_R/k}{SS_E/(N-k-1)}$$

当用解释的思想对回归模型进行检验，其目的是探讨每个自变量或自变量之间交互作用可以解释的变异占总变异的比例。可以使用下列公式：

$$R^2=\frac{\sum(Y_i-\overline{Y}_.)^2-\sum(Y_i-\hat{Y}_i)^2}{\sum(Y_i-\overline{Y}_.)^2}=\frac{SS_R}{SS_T}$$

其中，SS_R 是研究中预测因素（predictor）或自变量引起的变异。如果没有误差，预测因素或自变量引起的变异就是总变异，但这实际上是不可能的。R^2 表示预测因素或自变量引起的变异占总变异的比例。

本节简要描述了方差分析和回归分析的一般线性模型（general linear model），两种方法都包括建立矩阵。目前大多数计算机统计软件包在进行方差分析统计检验时都使用一般线性模型，因此熟悉这种途径对解释计算机的输出结果是很有帮助的。熟悉这种途径有助于帮助大家了解方差分析和回归分析的相似性，两种途径曾一度被研究者认为是截然不同的。在这一点上，一般线性模型特别吸引人，因为它把所有 ANOVA 实验设计模型和回归模型相融合。一般线性模型途径也提供了一个很好的方式使我们理解多变量研究，此外，还提供了对实验设计的基本概念的深入理解。

第四节　多重回归模型分析的实验举例

一、研究的问题与设计

这一节中，我们再通过一个研究举例来进一步说明使用解释功能时多重回归的特点。在一个关于儿童口语获得年龄的研究中（刘友谊，2006），研究者采用图片命名任务，通过回归分析检验获得年龄（AoA）、图片的名称一致性（NA）、概念一致性（CA）、概念熟悉性（Familiarity）、表

象一致性（Image）、视觉复杂性（V_Complexity）和图片名称的词频[log(1+fre)]等因素对图片命名反应时（RT_Harm）的影响，研究者特别对 AoA 与其他图片指标，尤其是与词频的关系感兴趣（见表 14-2）。

实验材料为 141 幅图片，来自斯诺德格拉斯（Snodgrass）的图库（Snodgrass & Vanderwart，1980）。实验中有七个自变量。AoA 指标是通过成人在一个九点量表上的主观评定获得的（见表 14-2）。得到的量表值可以分 1～2、2～3、3～4、4～5、5～9 五个水平，分别对应 0～3.5 岁、3.5～4.5 岁、4.5～5.5 岁、5.5～6.5 岁和 6.5 岁以后五个年龄段，它是一个分类变量。词频来自孙宏林等人（孙宏林等，1997）的现代汉语研究语料库，以每百万次出现的次数为指标，是一个连续变量。图片名称一致性、概念一致性、概念熟悉性、表象一致性、视觉复杂性指标来自于舒华等人（1989）的评定结果，均为连续变量。因变量是被试命名 141 幅图片的反应时。可以看出，在这个实验中有几个特点：自变量的数量比较多，为七个；自变量有分类和连续两种，且多数自变量是连续变量；141 幅图片来自一个现有的图库。数据处理有两种选择：一个是用方差分析处理数据，另一个选择是使用多重回归处理数据。用方差分析处理数据首先需要将连续变量转换为分类变量，缩减水平的数量。但可以想象，转换后很难保证各个处理结合条件下图片项目的数量是相同的，也就是说，由于实验中的变量、变量的水平众多，很难将 141 幅图片数量平衡地分配在几十个处理结合中。但是，如果不能在各个处理条件下平衡地分配项目，就无法满足方差分析中的 F 分布的要求，进一步影响显著性检验的可靠性。

表 14-2　141 幅图片 AoA 的分布情况

量表值	对应年龄段	图片/幅	图片累积/幅	所占百分比/%	累积百分比/%
1～2	0～3.5 岁	35	35	24.8	24.8
2～3	3.5～4.5 岁	36	71	25.2	50.4
3～4	4.5～5.5 岁	32	103	22.6	73
4～5	5.5～6.5 岁	23	126	16.4	89.4
5～9	6.5 岁以后	15	141	10.6	100

因此，研究者选择使用多重回归处理数据。数据分析的思路如下：首先对七个自变量（指标）与一个因变量进行相关分析，初步探讨各自变量

与图片命名反应时之间的关系，以及各变量之间的关系。接着进行完全多重回归分析，以图片命名反应时为因变量，其他七个变量为自变量，建立回归方程，探讨影响命名反应时的预测性指标。最后，为了更好地考察所关心变量 AoA 的独特的作用，采用层次回归（hierarchical regression）的方法，将感兴趣的变量放在最后一步进入方程，考察在排除了其他变量的贡献的情况下，该变量对回归方程的贡献。如果该变量仍然有明显贡献，那么可以作结论，即该变量确实具有其他变量所不能替代的独特的作用。

表 14-3 七个自变量及命名反应时的相关矩阵

	RT_Harm	AoA	log(1+fre)	NA	CA	Familiarity	Image	V_Complexity
RT_Harm	1.000							
AoA	0.272**	1.000						
log(1+fre)	−0.254**	−0.508**	1.000					
NA	−0.428**	0.042	0.022	1.000				
CA	−0.330**	0.098	0.044	0.548**	1.000			
Familiarity	−0.406**	−0.251**	0.231**	0.053	0.265**	1.000		
Image	−0.208*	0.072	0.087	0.131	0.146	0.089	1.000	
V_Complexity	0.122	0.131	0.045	0.066	−0.002	−0.416**	0.007	1.000

从表 14-3 的相关矩阵中可以看到，除视觉复杂性外，其他变量，如 AoA、词频、名称一致性、概念一致性、熟悉性、表象一致性等都与命名反应时存在比较高的相关，其中名称一致性跟反应时相关最高（$r=-0.428$，$p<0.01$），熟悉性次之（$r=-0.406$，$p<0.01$）。另外，七个变量之间也存在一定的相关。特别是研究者关心的一些变量之间存在复杂的相关，如 AoA 和词频之间的相关达到−0.508。变量之间存在相关，会

导致它们各自独立对因变量的贡献变得不清楚，这将是希望通过层次回归解决的问题。

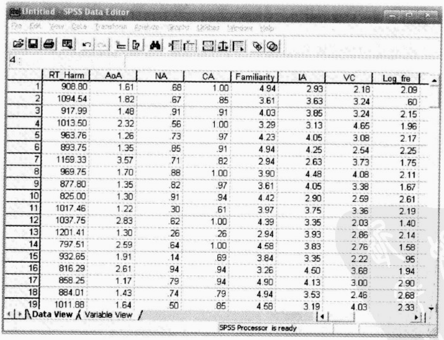
下面我们具体看看如何在 SPSS 中进行完全多重回归分析和层次回归分析。

二、多重回归模型分析的 SPSS 操作

（一）Enter 法

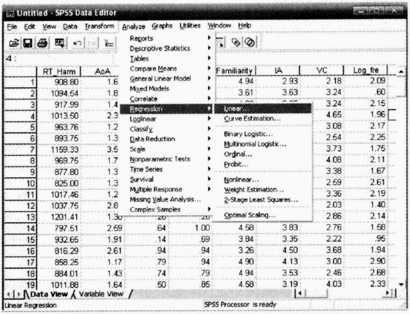
Enter 法是最简单一种估计回归方程式。它是将所选择的自变量全部进入建立的方程式中，是 SPSS 默认的方法。

研究者以图片命名反应时为因变量，AoA、词频、名称一致性、概念一致性、熟悉性、表象一致性和视觉复杂性七个因素为自变量，采用强行进入法（Enter 法）进行多重回归分析。所有变量在进行多重回归之前全部转化为标准分。

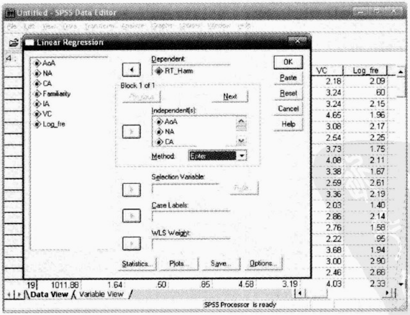


	RT_Harm	AoA	NA	CA	Familiarity	IA	VC	Log_fre
1	908.80	1.61	.68	1.00	4.94	2.93	2.18	2.09
2	1094.54	1.82	.67	.85	3.61	3.63	3.24	.60
3	917.99	1.48	.91	.91	4.03	3.85	3.24	2.15
4	1013.50	2.32	.56	1.00	3.29	3.13	4.65	1.96
5	963.76	1.26	.73	.97	4.23	4.05	3.08	2.17
6	893.75	1.35	.85	.91	4.94	4.25	2.54	2.25
7	1159.33	3.57	.71	.82	2.94	2.63	3.73	1.75
8	969.75	1.70	.88	1.00	3.90	4.48	4.08	2.11
9	877.80	1.35	.82	.97	3.61	4.05	3.38	1.67
10	825.00	1.30	.91	.94	4.42	2.90	2.59	2.61
11	1017.46	1.22	.30	.61	3.97	3.75	3.36	2.19
12	1037.75	2.83	.62	1.00	4.39	3.35	2.03	1.40
13	1201.41	1.30	.26	.26	2.94	3.93	2.86	2.14
14	797.51	2.59	.64	1.00	4.58	3.83	2.76	1.58
15	932.65	1.91	.14	.69	3.84	3.35	2.22	.95
16	816.29	2.61	.94	.94	3.26	4.50	3.68	1.94
17	858.25	1.17	.79	.94	4.90	4.13	3.00	2.90
18	884.01	1.43	.74	.79	4.94	3.53	2.46	2.68
19	1011.88	1.64	.50	.85	4.58	3.19	4.03	2.33

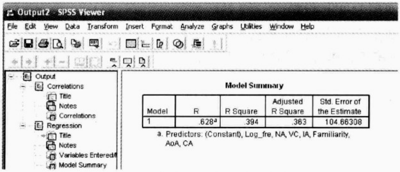
在 Regression 中选择 Linear 一项，然后弹出线性回归对话框。



在 Method 中选择 Enter 法，可以将自变量全部进入建立的回归方程中。



我们从图 14-1 中可以看到，使用 Enter 法，完全多重回归分析得出 $R^2=0.394$，表明七个变量能够解释命名反应时总变异的 39.4%。



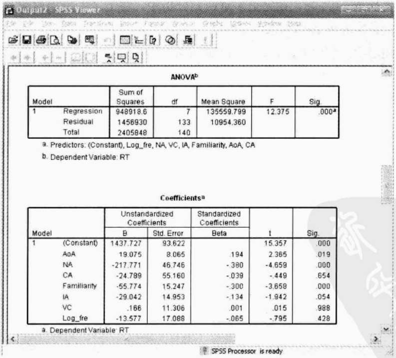
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.628[a]	.394	.363	104.66308

a. Predictors: (Constant), Log_fre, NA, VC, IA, Familiarity, AoA, CA

图 14-1　多重回归模型的输出结果

进一步的显著性检验表明，自变量可解释的因变量变异与误差变异相比是统计上显著的，$F(7, 134)=12.375$，$p=0.000$，表明回归方程是有意义的（见图 14-2 下半部）。研究结果与国外使用其他语言进行的 AoA



ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	948918.6	7	135559.799	12.375	.000[a]
	Residual	1456930	133	10954.360		
	Total	2405848	140			

a. Predictors: (Constant), Log_fre, NA, VC, IA, Familiarity, AoA, CA

b. Dependent Variable: RT

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1437.727	93.622		15.357	.000
	AoA	19.075	8.065	.194	2.365	.019
	NA	-217.771	46.746	-.380	-4.659	.000
	CA	-24.789	55.160	-.039	-.449	.654
	Familiarity	-55.774	15.247	-.300	-3.658	.000
	IA	-29.042	14.953	-.134	-1.942	.054
	VC	.166	11.306	.001	.015	.988
	Log_fre	-13.577	17.088	-.065	-.795	.428

a. Dependent Variable: RT

图 14-2　显著性检验的输出结果

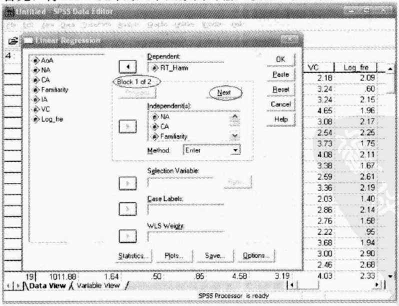
研究结果类似。系数分析还发现，在全部七种变量中，名称一致性、熟悉性和 AoA 三个变量对图片命名反应时有重要影响，但没有发现词频的作用（见图 14-2 上半部）。

由于 AoA 与词频有较高的相关，为了更好地考察所关心变量 AoA 的独特的作用，最后，研究者采用了层次回归（hierarchical regression）的方法。

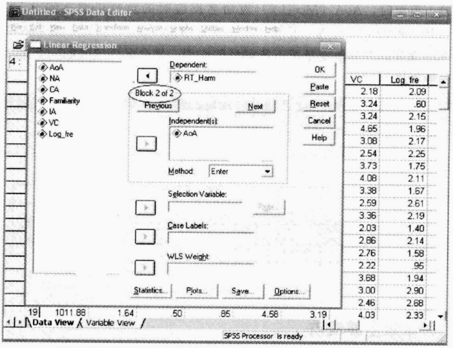
（二）层次回归

层次回归的基本思想是将感兴趣的变量放在最后一步进入方程，以考察在排除了其他变量的贡献的情况下，该变量对回归方程的贡献。如果该变量仍然有明显的贡献，那么我们可以作出该变量确实具有其他变量所不能替代的独特作用的结论。这种方法主要用于，当自变量之间有较高的相关，其中某一个自变量的独特贡献难以确定的情况。在此项研究中，研究者主要关心 AoA 在图片命名反应时中的独特作用。但是由于 AoA 和词频的相关达到－0.508，在完全多重回归分析中，将七个变量同时输入模型时，很难区分出 AoA 在图片命名反应时中的独特作用。因此，我们进一步使用层次回归的方法，共进行了两个步骤的分析。

首先，将 AoA 以外的六个因素同时输入模型，然后点 Next。



然后，把 AoA 输入自变量对话框中，并在 Statistics 选项中选择 R squared change 一项，可以观察到所关心的变量的独立作用。



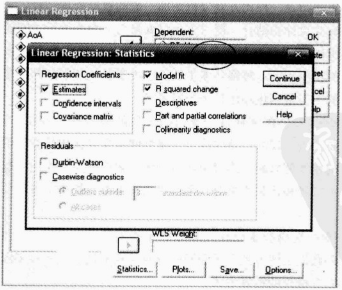
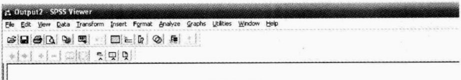


图 14-3 是模型输出的结果。层次回归分析结果发现，当将除 AoA 以外的六个因素先期输入模型，或控制了其他变量的影响，最后将 AoA 输入模型时，AoA（Model 2）对命名反应时的贡献仍然达到 2.5%（$\Delta R^2=0.025$），其贡献是统计上差异显著的（$p=0.019$）。结果表明 AoA 确实具有其他变量所不能解释的作用，是命名反应时的一个重要预测指标。



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.607[a]	.369	.341	106.44190	.369	13.058	6	134	.000
2	.628[b]	.394	.363	104.66308	.025	5.594	1	133	.019

a. Predictors: (Constant), Log_fre, NA, VC, IA, Familiarity, CA

b. Predictors: (Constant), Log_fre, NA, VC, IA, Familiarity, CA, AoA

图 14-3 层次回归分析的输出结果

本章主要观点

• 多重回归是一种比方差分析更加一般的数据处理途径，它可以与更加灵活的实验设计相结合，探讨多个自变量与一个因变量之间的关系。

• 回归模型分析有预测和解释两种功能：预测功能主要关心自变量对因变量的预测程度；解释功能主要关心由实验处理引起的因变量变化与误差引起的因变量变化相比是否差异显著。

• 当用回归分析解释自变量与因变量的关系时，与方差分析类似，研究者对每个自变量的主效应或自变量之间的交互作用的变异可以解释因变量变异的程度感兴趣。或者说，自变量主效应或交互作用引起的因变量变异与误差变异相比是否显著，其思想与在方差分析中平方和分解的思想是一致的。

• 一般线性模型途径将方差分析实验设计模型和回归模型相融合，有助于帮助我们了解方差分析和回归分析的相似性，也提供了深入理解实验设计的基本概念的很好方式。

• 层次回归的基本思想是将感兴趣的变量放在最后一步进入方程，以考察在排除了其他变量的贡献的情况下，该变量对回归方程的贡献。

思考题

1. 与多重回归模型分析相结合的实验设计的特点是什么？
2. 什么是回归分析的预测功能？什么是回归分析的解释功能？
3. 为什么可以说方差分析是回归分析的一个特例？
4. 层次回归分析主要可以回答什么样的问题？

第十五章
个案研究

前面的章节中介绍的实验研究的共同特点是使用多个被试的结果来分析心理机制，这类研究被称为组群研究。个案研究是通过对个案进行深入细致的探讨分析，从而比较深刻地认识一些特定现象的心理机制，因此也是心理学研究中的一种重要的方法。本章中，将结合认知神经心理学研究中语言障碍研究实例介绍个案研究的思想、方法及数据统计。

第一节　个案研究概述

前面的章节中介绍的研究的共同特点是，每种处理条件下选取多个被试，通过分析这些被试的平均值来考察实验处理的效应。这种以多个被试的结果来分析心理机制的研究被称为组群研究（group study)。使用多个被试的平均值来估计处理效应，其基本的思想是，许多心理机制、心理现象是人类共有的，处理效应平均数的分析可以使研究者更好地探讨人类心理的一般特点或典型机制。所以，组群研究一般适合以正常被试为对象的研究。在每种条件下使用多名被试的主要优点包括：可以使各个处理条件下的因变量平均值趋于稳定，可以计算组内的误差变异，可以减小实验中偶然因素带来的影响，因此得出的结果相对可靠，外部效度较高，结论能够被广泛地推广到其他个体等。所以，组群研究已经在心理学界得到普遍的应用。

然而，在有些情况下，组群研究的方法显示出一定的缺陷。首先，组群研究不适合对稀有现象的研究。研究者偶尔会发现一些稀有的个体，他们具有非常独特的心理特征，如果能够对其进行深入分析，将会得到非常有价值的成果。比如，一些病人因脑损伤使得其语义系统受到了损伤，但

这种损伤并不波及语义系统中的所有知识，而是仅局限在某个（或某些）特定的语义类别上，如动物类的损伤（见表 15-1）。表 15-1 中可以看出，患者 EA 在动物图形的命名上表现出明显的障碍（34%），而在其他类型图片的命名上基本保持完好。这种现象便是语义范畴特异性损伤（semantic category-specific deficits）。语义范畴特异性损伤的存在表明，人脑内的语义知识的存储或加工可能是分为不同类别的。语义范畴是语义系统的一个重要的组织维度，由此提高了研究者对语义系统的内部结构进行深入细致探讨的热情。但这种现象十分罕见，发生率极低，几乎很难达到组群研究的被试量标准，所以我们研究它就不能使用常规的组群研究。

表 15-1　患者 EA 在命名不同语义类别图形的正确率

语义类别	正确率
动物	34%
人体器官	92%
衣物	100%
水果	100%
家具	100%
厨具	94%
乐器	80%
工具	83%
蔬菜	100%
交通工具	93%
其他	87%

（引自 Caramazza & Shelton，1998）

其次，有时组群研究也不适合研究个体变异较大的现象。即使有些研究的被试量似乎能够达到组群研究的要求，但仔细分析后发现，这些被试的个体差异极大，基于这些个体数据得出的平均数可能并不代表任何典型的心理机制。也就是说，如果用被试群体的平均值取代每名被试的表现，其实会出现很大偏差。例如，为了探讨汉语阅读障碍儿童（dyslexic children）的障碍机制，吴思娜等人（吴思娜等，2004）经过大量筛选得到了

15名阅读障碍儿童。但进一步分析后发现，这些儿童间的个体差异非常大。尽管所有障碍儿童的语素意识测验得分均低于正常水平，但是他们在语音意识、同音判断和语义相关判断三个认知测验上出现了明显的分离，表明这些儿童还可细分为不同的阅读障碍亚类型（见图15-1）。在这种情况下，如果将他们的数据平均起来处理，就很可能掩盖其真正的阅读障碍机制。很显然，如果我们要对上述现象进行深入细致的探讨，就不能采用传统意义上的组群研究，而只能利用要求被试量较少，甚至只有一名被试的研究，这就是通常所说的个案研究（case study）。

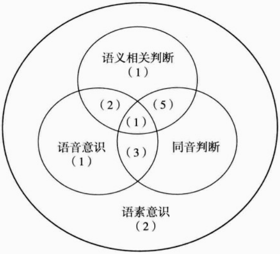


图15-1　15名阅读障碍儿童在四种缺陷中的分布情况（括号内的数字是儿童数）
（引自吴思娜等，2004）

个案研究不仅对描述一些稀少事件、个体差异大的心理现象非常重要，而且也对建立理论模型发挥了积极作用。比如，语言运动理论曾经认为语言信号译码能力依赖于听者说的能力，如果一个人不能说，他就不能理解别人的话。然而，伦贝里（Lenneberg，1962）描述了一个8岁男孩，他的语言运动技能受损，却可以清楚地理解别人说的话。这一个案强有力地反驳了语言运动理论的基础。再比如，精神分析理论来自于弗洛伊德（S. Freud）本人对个案的观察与思考，儿童发生发展理论是皮亚杰（J. Piaget）利用少量个体建立起来的，人类记忆的遗忘曲线是艾宾浩斯（H. Ebbinghaus）以自己为被试绘制出来的，条件反射是巴甫洛夫

(И. П. Павлов) 利用仅有的几条狗发现的。凡此种种，不胜枚举。这些重要的心理学研究成果主要是通过个案研究的方法获得的。

时至今日，个案研究已经受到了学者们越来越多的重视，在认知心理学、社会心理学、医学心理学等领域都得到广泛应用，研究成果不断涌现，甚至有些重要的国际心理学杂志开始主要刊载这方面的成果，如 *Cognitive Neuropsychology* 和 *Neurocase* 等。可以说，个案研究是组群研究的很好补充，它不仅丰富了心理学的研究手段，而且也为心理学理论模型的建立作出了非常大的贡献，许多重大的心理学成果就是利用这种方法得来的。

第二节　个案研究的思想和方法

一、个案研究的思想

虽然从诞生之日起，个案研究就得到了一些学者的认可，但同时也被很多人质疑。被质疑的一个主要方面是，个案研究使用的被试人数较少，这样得出的结论是否有代表性，能否将其结论推广到其他个体，能否通过个案研究获得一般性的结论。为了澄清这些疑问，下面不妨让我们以认知神经心理学（cognitive neuropsychology）的研究为例，来简要说明个案研究的基本思想。

在认知神经心理学研究中，个案研究是一种非常普遍的研究方法。它主要是基于对脑组织受损患者的细致测试，确定患者受损和保留的认知环节，从而对正常人的心理机制作出相对准确的探讨。举例来说，在临床上，一些脑损伤患者经常出现命名障碍，他们一般表现为在命名物体或图形时有困难，出现大量的命名错误。然而，这些命名障碍患者损伤的认知水平却可能是不同的。研究发现，如果损伤在语义环节，病人在命名图片时只出现语义错误（如将图片 chair 命名为 table）和无关错误。而对于语音输出词典选择性损伤的病人，他们命名图片时除了犯语义错误外，还可能包括语音相似错误（如将图片 cat 命名为 mat）和非词错误。这些来自特异性损伤患者的研究结果支持在人的言语产生中存在独立的语义系统和语音输出词典的理论模型（见图 15-2）。然而，由于正常人命名物体、图

形的过程非常自动化，无论用反应时或错误率的方法都很难观测到语言产生的细微的加工环节，以得到非常有力的证据支持独立的语义系统和语音输出词典的理论模型。

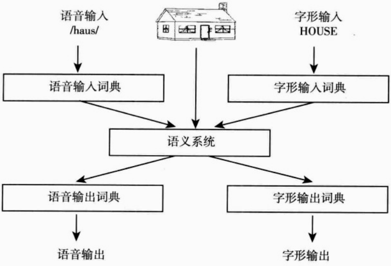


图 15-2 词典系统的功能结构简图

假设的建立对认知神经心理学个案研究尤为重要，研究者需要在假设的指引下，发展出适宜的测试方法，并对数据进行合理分析，正确推论，最后上升到建立理论或模型。认知神经心理学个案研究的前提假设之一就是功能结构无个体差异（uniformity of functional architecture across people），即对于不同的个体来说，他们基本的功能结构（内部的心理机制）是相同的。一名个体基本的功能结构可以类推到其他个体，所以一项个案研究的结论能够在一定程度上进行推广。例如，目前的研究表明，人类的心理词典系统（人脑中加工语言信息的系统）至少包含五个相互关联的部分：语音输入词典、字形输入词典、语义系统、语音输出词典和字形输出词典（见图 15-2）。听觉词汇理解时，听觉刺激首先激活语音输入词典中的语音输入信息，这些信息然后激活语义系统中的语义信息，从而实现对词汇的理解。口语词汇产生时，首先激活语义信息，然后激活相应的语音输出词典中的语音输出信息，这些信息得到进一步加工后，直到最终发出目标语音。视觉词汇理解和词形产生的过程也经历类似的过程。一般情

况下，每个人的心理词典系统的结构是大体相同的，它不因种族、性别等的不同而存在差异。也就是说，如果我们能够通过对少数特殊被试的深入测试弄清楚该系统的结构和加工，就可能更好地认识人类的心理词典系统。

在有关认知神经心理学的个案研究中，一般需要进行严谨的实验设计，在每个测验中测试大量的项目，并对数据进行推论统计检验。其实，个案研究暗含的前提假设与组群研究的前提假设是类似的。组群研究的假设是，通过测试大量的被试，以及推论统计检验，使研究结论能够推广到更一般的情况。个案研究则是通过对一个被试测试大量的项目，以及推论统计检验，使研究结论能够推广到更一般的情况。从建立一个普遍的心理学模型或理论的角度，组群研究和个案研究也是互补的。如果一个模型或理论只能对组群研究的平均值加以解释，而对个案的行为不能合理说明，表明这个理论还存在缺陷，还需要进一步完善。一个比较科学的心理学理论应该既能解释组群研究的数据，也能解释个案研究的数据，不仅能解释正常人的行为，也能解释病人的行为。

二、个案研究的方法

个案研究具有自身的特殊性，人们在长期研究经验积累的基础上为它发展出了一些独特方法。下面仅介绍几种常用的方法：分离（dissociation）和相关（association）的方法、个案—对照组（case-control group）的方法。

（一）分离和相关的方法

分离和相关的方法在认知神经心理学中运用得比较普遍。在整个研究过程中，能够准确细致地确定患者的障碍环节和保留环节是这种研究方法的关键。研究者往往要对患者进行多种任务的施测，通过比较患者在不同任务中的作业成绩，着重分析患者表现出障碍的任务与其他任务之间的关系，以便确定患者的具体障碍所在，并上升到理论高度。这种关系被分为相关和分离两种，而分离又分为单分离（single dissociation）和双分离（double dissociation）。

1. 单分离

单分离的基本含义是在同一类控制变量下要求患者完成两种不同任

务，如果患者仅在其中一种任务上表现出障碍，或在其中一种任务上的障碍程度比在另一种任务上的障碍程度严重，则这两种任务之间表现为单分离。具体地说，当认为某患者在两个特定任务之间存在单分离时，首先要保证患者病前在这两种任务中均保持正常水平，而且两种任务间的材料因素（如熟悉性、出现频率等）和被试因素（智力、文化程度、年龄、病情稳定性等）也得到了控制，同时还要排除患者的障碍是由于认知水平之外的损伤造成（如不能书写是由于肢体瘫痪引起）的可能性，从而可以推断患者表现出的障碍可能主要是认知水平的，而与外周因素相关较小。在严格控制各种无关因素的情况下，如果患者仍表现为在一种任务上的障碍程度比在另一种任务上明显严重，便可以认为这两种任务之间具有单分离。出现单分离的最可能原因是两种任务的认知机制具有一定的独立性，作业成绩差的认知机制出现了更严重的障碍。所以，利用该方法有助于把认知环节进一步区分与细化，有助于区分两个独立的加工环节，使得人们对心理机制的认识更细致、更深入。

我们举一个例子来说明单分离方法的使用。例如，患者 HW (Caramazza & Hillis，1991）病前是一位语言功能正常的售货员，后因左脑顶叶中风出现了严重的口语产生障碍，包括朗读、图形命名均不能达到正常水平。通过皮博迪图词测验（Peabody Picture Vocabulary Test，简称 PPVT）的测查表明，她对词汇的听觉和视觉理解能力都正常。同时，如果不考虑少许拼写错误，她的书写结果也近乎正常。此外，她的短时记忆、非词典系统（如视力、听力等）也在正常范围。研究者初步得出结论，HW 的障碍仅局限在语音输出通道上，而且这种障碍的起因是来自词典系统本身，与非词典因素关联较小（参见图 15-2)。在初步的测试中，研究者还发现 HW 在口语产生时表现为动词产生比较困难，而名词产生相对容易，即可能存在动词特异性损伤（verb specific deficits)。为了进一步考察这一非常有意义的词类选择性损伤形式，研究者设计了如下更细致的测验。为了证实 HW 在口语产生时是否确实存在动词、名词差异，他们让 HW 完成图形命名任务。具体的做法是，首先选出 60 幅彩色图形，动词图形（也就是动作图形，如走）和名词图形（也就是实物图形，如狗）各 30 幅，并对动词图形和名词图形的名称在词长、词频等多

种因素上进行了匹配。然后要求 HW 对所有的动词和名词图形进行命名任务，即说出每幅图形的名称。另外，为了比较 HW 在口语产生与书写任务上的差异，研究者也让 HW 对上述图形完成图形写名任务，要求她写出每幅图的名称。结果表明，在 30 幅名词图形中，HW 正确命名 16 幅，正确书写 29 幅。然而，在 30 幅动词图形中，她只正确命名 6 幅，正确书写 29 幅（见表 15-2）。结果中可以看到 HW 命名动词图形的成绩（20%）明显差于命名名词图形（53%）。这就是一个典型的单分离研究案例。研究者通过单分离发现了 HW 存在动词特异性障碍，而且发现她的动词特异性障碍仅仅局限于语音输出通道上。

表 15-2　HW 在两任务中名词、动词的正确率

	图形命名	图形写名
名词	53%（16/30）	97%（29/30）
动词	20%（6/30）	97%（29/30）
合计	37%（22/60）	97%（58/60）

然而，人们也经常质疑单分离现象的解释，认为患者在不同任务间作业成绩的差异可能不是来自认知环节之间的分离，而是由于一些外在因素（如材料性质不同，任务难度存在差异等）导致的结果。例如，对于 HW 这个例子来说，也可能因为动词材料本身比名词材料更抽象、更低频，因而导致 HW 在动词命名上成绩较差。这对单分离方法得出的结论提出了极大的挑战，然而单分离方法的局限又不容易直接对疑问加以反驳。另外一种思路，如果我们能够发现另外一类患者，对他们进行相同任务的施测，在严格控制了其他影响因素后，他们表现出与前一类患者截然相反的结果模式，那么就可以消除可能由于材料等人为因素带来的影响，从而确认分离是基于认知功能的，而不是人为因素造成的。这种在两种任务间交互补充的分离模式，即为双分离。

2. 双分离

双分离的基本含义为在同一类控制变量下要求患者完成两种不同的任务。第一位病人仅在其中一种任务上表现出障碍，或在其中一种任务上表现的障碍程度比在另一种任务上的障碍程度严重，而第二位病人在两种任

务之间的障碍模式与第一位截然相反，则这两种任务之间表现为双分离。

我们来看一个双分离的例子，例子中涉及两位患者：患者 HW 和患者 EBA（Hillis & Caramazza，1995）。研究者给患者 EBA 施测与患者 HW 完全相同的测试材料，发现 EBA 也存在口语产生障碍。但是进一步的分析发现，EBA 的结果是，在图形命名时，名词的正确率为 10%（3/30），动词的正确率为 70%（21/30），前者明显低于后者（$\chi^2_{(1)}=22.500$，$p<0.0001$），EBA 表现为名词特异性障碍。这样，她与 HW 的动词特异性障碍形成了双分离模式（见表 15-3）。EBA 的结果可以有力地反驳“患者 HW 在动词命名上的困难是因为动词材料本身比名词材料更抽象、更低频”的说法。基于这种双分离模式，可以进一步说明，动词命名和名词命名之间的分离是基于认知水平的分离，而不是材料难度等外部因素造成的结果。

表 15-3　患者 HW 和 EBA 在图形命名中动词、名词的正确率

	HW	EBA
名词	53%（16/30）	10%（3/30）
动词	20%（6/30）	70%（21/30）

让我们再举另外一个双分离的例子，患者 AS 和 IFA（Caramazza，et al.，2000）都存在复述困难。然而他们在复述任务中，构成了元音和辅音的双分离：患者 AS 表现为复述元音差，复述辅音较好；患者 IFA 表现为复述元音较好，复述辅音差。例如，如果把他们复述的各个词汇分解为单个的元音与辅音来记分，那么 AS 对元音和辅音的错误率分别为 26.9%（737/2 736）和 9.3%（318/3 434），前者明显高于后者（$\chi^2_{(1)}=335.668$，$p<0.0001$）；而 IFA 对元音和辅音的错误率分别为 5.3%（181/3 397）和 28.2%（1 173/4 159），前者明显低于后者（$\chi^2_{(1)}=665.232$，$p<0.0001$）。另外，如果让他们复述同一套具有 CVCVCVCV 结构的词（“C”和“V”分别表示这个词汇中该位置是辅音和元音），结果也同样发现，在整个词汇内部，患者 AS 复述元音很差，复述辅音较好，而患者 IFA 与其正好相反（见图 15-3）。由此说明，在复述这个心理加工过程中，元音、辅音确实可能在表征或加工的某个环节具有相对独立性。

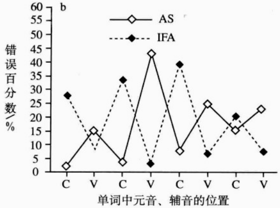


图 15-3　AS 和 IFA 复述元音（V）和辅音（C）的错误率

（引自 Caramazza，et al.，2000）

此外，更有趣的是，双分离现象还可能发生在同一个患者身上。例如，患者 KSR（Rapp & Caramazza，2002）对相同的测验材料分别完成了口语产生和书写产生任务（图形命名、范畴流畅性、句子产生）。结果发现，患者 KSR 在口语产生时表现为名词特异性障碍，而在书写产生时却表现为动词特异性障碍（见表 15-4）。这个案例进一步说明在词典系统中，动词和名词信息是具有相对独立性的。

表 15-4　患者 KSR 在口语、书写产生中的调查结果

	口语产生	书写产生
动词	×	××
名词	××	×

注：×表示轻微损伤；××表示严重损伤。

从前面的分析可以看出，与单分离方法类似，双分离方法有助于区分两个相对独立的认知加工环节，可以帮助研究者进一步区分和细化认知环节。由于双分离方法利用双被试甚至是一个被试本身作控制，消除了可能由于材料等人为因素带来的影响，因此它要比单分离方法的结果更可靠、更有说服力。这种方法受到了研究者的高度重视。

3. 相关

分离的方法主要是通过考察两种任务的损伤不均衡，来帮助人们把一

些难以区分开的认知功能分离出来。但是在个案研究中，仅采用这种方法还远远不够。一般来说，分离的方法只能大体地告诉人们两个认知功能存在分离，但还不能具体地告诉我们，这种分离到底发生在哪个特定的环节。比如，通过研究患者 HW 和患者 EBA，我们得知在语音输出通道上，动词和名词信息可能是相对分离的，但这种分离到底具体发生在哪个环节，是加工句法信息的环节还是加工语音信息的环节，是存储机制还是加工过程受损等，还不很清楚。为了弄清这些问题，需要借助相关的方法加以解决。

相关的基本含义是在同一类控制变量下要求患者完成两种不同任务，如果患者在两种任务上均表现出障碍，则这两种任务之间表现为相关。出现相关的最可能原因是完成两种任务的认知功能具有一定的共享性，患者的这个共享机制出现了损伤。所以，利用相关的方法有助于找出两个功能之间共同的认知环节。

我们举例来说明相关方法。研究者要求患者 JJ 对一套动物和非动物图形进行口语命名。研究者事先对两类图形名称的词长、词频等指标进行了细致的匹配。结果发现，患者 JJ 对动物图形的命名正确率（42/46）远远高于对非动物图形的命名正确率（20/98）（$\chi^2_{(1)}=64.178$，$p<0.0001$），表明他在图形命名时存在动物类的选择性损伤（Caramazza & Hillis，1991）。那么患者 JJ 的动物类损伤到底发生在语言加工系统的哪个环节呢？由于图形命名包含的主要环节包括视觉输入系统、语义系统、语音输出系统等（参见图 15-2），所以研究者下一步的基本思路为，如果损伤发生在语义系统，那么患者应该在需要语义系统参与的所有任务中，如图形写名、听觉图—词核证（听觉理解任务）、视觉图—词核证（视觉理解任务），都应该表现出动物类损伤。也就是说，如果患者 JJ 在所有这些任务上表现为动物类损伤，表明这些任务之间出现相关，患者的损伤是发生在完成各种任务所共同需要的认知环节—语义系统。但是，如果损伤发生在语音输出系统，则患者 JJ 应该在不包括语音输出的任务中，如图形写名、听觉图—词核证、视觉图—词核证，不表现为动物类的损伤，即语音输出任务与非语音输出任务在动物—非动物类项目上出现分离，而不是相关。研究者对患者 JJ 在上述任务上进行了测试，实际的测验结果表明，患者

JJ 在需要语义系统参与的任务上均表现为动物类的选择性损伤（见表 15-5）。由此说明，JJ 的动物类选择性损伤发生在语义系统。

表 15-5　患者 JJ 在不同通道任务上动物与非动物项目的正确率

语义类别	图形命名	图形写名	听觉图—词核证	视觉图—词核证
动物	91.3%	70.0%	91.3%	97.8%
非动物	20.4%	15.3%	60.2%	42.9%

实际上，从事个案研究时，只有把分离和相关的方法结合起来灵活应用，才可能对个案的具体模式作出比较清楚准确的认识。如果把两种方法割裂开来使用，往往会对结果的解释出现偏颇，甚至可能会得出错误结论。

分离和相关的方法具有的明显特点是，主要借助患者自身在完成不同任务之间，或通过患者与患者完成任务的相互比较，来推知患者某种认知功能的损伤或保留情况。还有一种方法也同样能达到类似的研究效果，那就是通过患者与正常对照组的相互比较，来确认患者的障碍模式，这就是下面要介绍的个案—对照组的方法。

（二）个案—对照组的方法

个案研究中，研究者感兴趣的问题往往是个案的一些相对异常的心理功能，探讨它们的起因、程度、发展等内部机制。个案—对照组的研究方法的基本思想和方法是，为个案设立一个正常对照组，个案和对照组除了在研究者要探讨的功能上可能存在差异外，在年龄、性别、受教育程度等其他无关因素方面尽量保持一致。这样，研究者可以通过个案和对照组被试在完成各种任务上的比较，揭示个案的异常的心理机制。在这类研究中，对照组选取是否恰当是至关重要的，因此选取过程一定要小心谨慎。下面我们举例来说明。

例如，在一项发展性阅读障碍研究中，栾辉等人（2002）对一名小学四年级的阅读障碍儿童 J 进行了深入的探讨。初步访谈发现，儿童 J 没有神经、行为、情绪的障碍，感觉、动作能力正常，但他的语言能力较差，包括口语表达、阅读和理解。进一步测试表明，J 的智力正常，言语智商为 102，但汉语识字量测验结果表明 J 的识字量很差，比同年级的儿童约

低 1.7 个年级。为深入认识儿童 J 的障碍特点，研究者选择了 5 名正常儿童作为 J 的生理年龄对照组。这 5 名儿童与 J 来自相同的学校、班级，年龄、智力与 J 相类似。儿童 J 及其对照组的生理年龄和智商的详细资料见表 15-6。

表 15-6 个案与对照组的描述性资料

	儿童 J	生理年龄对照组（n=5）
生理年龄	10 岁 8 个月	10 岁 8 个月（10 岁 1 个月至 10 岁 11 个月）
RAVEN	>75%	>75%（50%～95%）
WISC-V	102	102（93～110）

研究者对 J 和对照组进行了一系列认知能力测验，尤其对他的语音意识和汉字命名进行了细致的测验。结果发现，在语音意识测验中，J 的基本辨音能力正常，也能够完成一些基本的语音任务，如同音节判断、押韵判断等，能够完成在音节、韵律水平上的语音操作。但 J 在精细的语音加工，如音位删除、声调辨别任务上的正确率明显低于对照组儿童，表现出语音意识的缺陷（见表 15-7）。在这个研究中，通过与正常对照组的比较，揭示了儿童 J 存在的语音障碍的细节与特点。这些细节与特点对于以后的矫治和训练是十分重要的。

表 15-7 儿童 J 与对照组语音意识测验结果

类别	任务	方式	项目数	J		对照组		差别
				正确数	正确率	正确数	正确率（最小值～最大值）	J—对照组
辨音	听写拼音	单音	12	9	0.75	10	0.87(0.75～1)	−0.12
		在词中	12	10	0.83	11	0.90(0.83～1)	−0.07
听觉记忆	语音重复		10	7	0.7	9	0.92(0.90～1)	−0.22
音节意识	判断同音节	真词	36	35	0.97	35	0.98(0.94～1)	−0.01
		假词	36	35	0.97	35	0.98(0.97～1)	−0.01
韵律意识	押韵判断	视觉呈现	36	32	0.89	36	0.99(0.94～1)	−0.1
音位意识	音位删除	听觉呈现	16	6	0.38	14	0.93(0.75～1)	−0.55*

续表

类别	任务	方式	项目数	J			对照组	差别
				正确数	正确率	正确数	正确率（最小值～最大值）	J－对照组
挑异音	听觉呈现	声母	12	6	0.5	10	0.89(0.75～0.92)	－0.39
		韵母	12	6	0.5	9	0.87(0.58～0.83)	－0.37
		声调	8	2	0.25	7	0.72(0.88～1)	－0.47*
	视觉呈现	声母	12	11	0.92	12	1(1～1)	－0.08
		韵母	12	11	0.92	10	0.83(0.67～0.92)	0.09
		声调	8	3	0.38	8	0.98(0.88～1)	－0.6*

个案—对照组方法更多适用于一些典型的个案研究。在一些情况下，我们还可以针对包含少量被试的个案小组来匹配对照组，即个案（障碍）组—对照组方法。在有关老化、发展性障碍等领域的研究中，个案组和对照组之间的比较也是一种常用的方法。我们再举一个例子来说明。在一个有关汉语阅读障碍儿童的研究中，吴思娜等人（2004）采用了障碍组—对照组的方法。他们首先选出同时满足以下两个标准的儿童作为阅读障碍儿童：（1）在标准化汉字识字量测验中至少低于正常 1.5 个年级；（2）命名组词测验中至少低于正常儿童的 20%。这样，从测试的 211 名儿童中得到 15 名阅读障碍儿童。为了深入研究这 15 名儿童的障碍模式，研究者又从相同的学校、班级选出了 15 名正常儿童作为他们的对照组（其中一名学生中途转学，剩余 14 名作为对照组），两组儿童之间除了在识字量和命名组词能力上存在明显差异外，其他方面基本同质，如智力（依据瑞文推理测验）和年龄，详情见表 15-8。

表 15-8 阅读障碍儿童组与对照组的匹配情况

	障碍组（$n=15$）	对照组（$n=14$）
年龄/岁	11.6	11.7
瑞文推理测验得分	69%	72%
识字量/个	2 070.7	2 838.0**
命名组词的正确率	72%	94%**

研究者对障碍组和对照组在各个测验上的平均数之间的差异进行了分析（见表 15-9）。结果发现，障碍组儿童在语音意识和语素意识测验上的正确率都低于对照组儿童的正确率。障碍组儿童的语音通达和语义相关判断的反应时比对照组儿童更长，错误率更高。从描述统计结果可以看出，障碍组和对照组儿童在语素意识、语音意识、词典通达速度、语义相关判断的速度和错误率上都存在着差异。

表 15-9　两组被试各测验的描述统计

	障碍组	对照组
语音测验（正确率）	0.83（0.1）	0.92（0.06）
语素测验（正确率）	0.46（0.14）	0.77（0.05）
同音判断“是”反应（毫秒）	990（284）	728（117）
“否”反应（毫秒）	1 106（402）	861（153）
错误率	0.1（0.09）	0.1（0.08）
语义相关判断（毫秒）	1 377（380）	1 043（271）
错误率	0.26（0.15）	0.12（0.08）

注：括号内的数字为标准差。

一般来说，个案（障碍）组—对照组方法更关心的是，与正常群体相比，特殊群体的一些共性的特点，而在一定程度上忽略特殊群体中的个体差异。

三、个案研究的统计检验

在现代心理学研究中，个案研究不仅仅是简单地收集个别被试的数据，进行描述统计。要通过个案研究获得一般性的科学结论，也需要推论统计，而且推论统计在个案研究中是非常重要的，也是必需的。个案研究中如何进行统计检验？它的检验方法与一般组群研究的检验方法有何差别？为了说明这些问题，让我们先来了解一下个案研究中收集数据有什么特点，然后再具体介绍它的几种主要的检验数据的方法。

总体来看，个案研究中所收集的数据主要有三个特点。（1）在大量个案研究中，研究者收集的数据是被试行为的正确数、错误数等，这些数据

一般是计数数据（如 5 个、16 个……）。(2) 在许多个案研究中，只有一两个被试，虽然研究中可以设计不同的实验处理条件，但一个被试的行为数据是很难得到正态或接近正态分布的。(3) 在许多情况下研究者使用个案（组）—对照组方法，这时需要将一个被试的行为数据和一个小组被试的行为数据相比较，或将一个实验组与对照组的数据进行对比。

可见，个案研究的数据有明显不同于一般组群研究的数据的特点。在很多研究中，组群研究的数据多是连续型随机变量，而且我们假定这些随机变量大体服从正态分布。为此，个案研究的检验方法与多个被试平均数的方法存在差别。目前人们发展出了多种检验个案研究结果的方法，其中最常用的是卡方检验（chi-square test）和 t 检验。下面就着重介绍一下这两种方法。

（一）卡方检验

卡方检验是处理离散数据的一种检验方法，属于一种非参数检验。它主要有两种用途：拟合度检验和独立性检验。前者是解决一个因素内两项或多项分类的实测值（实际测得的观察值）与有关总体的理论值（理论上的预期值）是否一致的问题。它适用于一个因素多项分类的计数数据。而后者是检验两个因素是否有独立性或有无相互关联。它适用于两个因素多项分类数据的检验，一般采用列联表来记录数据。如果两个因素分别含有 r 个和 c 个分类项目，则可以建立一个 $r\times c$ 的列联表（如 2×2，3×5，4×4）。其中 2×2 的列联表是最简单也是最常用的一种形式。由于它由四个数据格组成，所以人们通常也把它称为四格表。下面我们将通过举例的方式对这几种检验逐一介绍。在此之前先来了解一下卡方值的计算公式。

卡方检验主要考察实测值和理论值的吻合程度。实测值和理论值越吻合，则表明个体的实际表现与理论预期假设越接近。这种吻合程度可以通过卡方值的大小反映出来。卡方值的计算公式如下：

$$\chi^2=\sum\frac{(f_o-f_t)^2}{f_t}$$

其中，f_o 和 f_t 分别表示实测值和理论值。

另外，卡方检验过程中也需要求得自由度。具体做法是，在拟合度检

验中，如果一个因素分为 k 个项目，则它的自由度是 $df=k-1$。在 $r\times c$ 列联表检验中，自由度是 $df=(r-1)(c-1)$。在四格表检验中，自由度是 $df=1$。下面我们更加详细地举例说明。

1. 拟合度检验

我们以前面提到的关于患者 HW 的研究为例。研究者发现 HW 在口语产生时表现为名词产生比较困难。HW 在命名 30 个名词图形的名称时，正确命名了 16 个图形名称，错误命名数目是 14 个。而与 HW 的年龄和受教育程度相似的对照组在命名这些名词图形的名称时，正确数与错误数的比例约为 9∶1。现在我们想检验一下 HW 的这种命名表现与其对照组是否一致，可采取如下步骤。

第一步，建立假设。

提出如下的虚无假设（H_0）和备择假设（H_1）：

H_0：HW 的命名正确与错误项目的比率为 9∶1；

H_1：HW 的命名正确与错误项目的比率不是 9∶1。

第二步，计算卡方值。

根据虚无假设 H_0，正确数与错误数的理论值分别为：

$f_{t正确}=30\times9\div10=27$，$f_{t错误}=30\times1\div10=3$

根据卡方值的计算公式，求得卡方值：

$$\chi^2=\frac{(16-27)^2}{27}+\frac{(14-3)^2}{3}=44.815$$

另外，它的自由度为 $df=2-1=1$。

第三步，统计检验。

根据 χ^2 值表查得，当自由度为 1，显著性水平 p 为 0.05 和 0.01 时，χ^2 值分别为 3.84 和 6.63（见 χ^2 值表）。然后再将实际计算出的 χ^2 值与之相比较。由于 $\chi^2=44.8>6.63$，则 $p<0.01$。按照统计决断规则，应该在 0.01 显著性水平上拒绝虚无假设 H_0 而接受备择假设 H_1。所以我们可得到的结论是，患者 HW 在命名名词图形时的命名正确项目数与错误项目数比例与其对照组是不一致的。

这里需要说明的是，如果有一个或多个理论值小于 5，则需要进行统计校正。

2. 独立性检验

（1）列联表检验。列联表检验适用于两个因素多项分类数据的检验，我们举例来说明其检验过程。例如，有一位传导性失语症患者 FS，他存在明显的复述障碍。为了探讨 FS 在复述过程中是否存在词类效应，我们要求他复述了 160 个名词、166 个动词、135 个形容词和 100 个副词。分析结果时，我们把他的复述成绩按正误分为三种类型，分别为完全正确、部分正确、完全错误。结果如表 15-10 所示。

表 15-10　FS 复述各类词的正误成绩

正误	词类				合计
	名词	动词	形容词	副词	
完全正确	22	26	74	85	207
部分正确	56	53	52	59	220
完全错误	82	87	83	41	293
合计	160	166	209	185	720

如果我们现在想知道，FS 在复述上述四类词时，在三种正误类型分布上是否有差异，则相当于进行一个 4（词类：名词，动词，形容词，副词）×3（正误：完全正确，部分正确，完全错误）的列联表检验，可采用如下步骤。

第一步，建立假设。

H_0：FS 复述四种词类时，在三种正误类型上无差异；

H_1：FS 复述四种词类时，在三种正误类型上有差异。

第二步，计算卡方值。

从表 15-11 中可以看出，我们共得到了 12 个实测值（不包括合计栏目的数值）。按理论假设，每个实测值都对应一个理论值，它的算法是：

每一观察值的理论值＝所在横行项目总和×所在纵列项目总和÷所有项目总和

例如，名词完全正确的观察值为 22，它的理论值＝207×160÷720＝46。这样便可以依次求出每个实测值的理论值，结果见表 15-11，表中括号内的数值就是理论值。

表 15-11 FS复述成绩的实测值和理论值（括号内）

正误	词类				合计
	名词	动词	形容词	副词	
完全正确	22 (46.000)	26 (47.725)	74 (60.088)	85 (53.188)	207
部分正确	56 (48.890)	53 (50.722)	52 (63.861)	59 (56.528)	220
完全错误	82 (65.111)	87 (67.553)	83 (85.051)	41 (75.285)	293
合计	160	166	209	185	720

然后根据卡方检验的公式，求出卡方值：

$$\chi^2=\frac{(22-46)^2}{46}+\frac{(26-47.25)^2}{47.725}+\cdots+\frac{(41-75.285)^2}{75.285}=73.75$$

自由度为 $df=(4-1)(3-1)=6$。

第三步，统计检验。

根据 χ^2 值表查得自由度为 6，显著性水平 p 为 0.05 和 0.01 时的 χ^2 值分别为 12.59 和 16.81。由于得到的 $\chi^2=73.75>16.81$，$p<0.01$。所以，应该在 0.01 显著性水平上拒绝虚无假设而接受备择假设，最终得到的结论是，FS 复述四种类型的词时，在三种正误类型的分布上存在显著差异。

(2) 四格表的检验。四格表是 2×2 的列联表，是列联表的一个特例，它的检验方法与列联表完全相同。

我们再以患者 HW 为例，她在命名图形时，在 30 幅动词图形中正确命名了 6 个，错误命名了 24 个。而在 30 幅名词图形中，正确命名了 16 个，错误命名了 14 个（见表 15-12）。如果我们想知道她命名动词图形和名词图形的正确和错误项目的分布模式上有无差异，也就是词类是否会影响她的命名表现，这就相当于要检验一个四格表：2（词类：动词、名词）×2（正误：正确、错误）。

首先，根据公式求得各实测值的理论值，结果见表 15-12，括号内的数据就是相应的理论值。随后根据表 15-12 中的结果，再求出卡方值：

$$\chi^2=\frac{(16-11)^2}{11}+\frac{(6-11)^2}{11}+\frac{(14-19)^2}{19}+\frac{(24-19)^2}{19}=7.177$$

它的自由度为 $df=1$。

表 15-12　HW 的图形命名成绩的实测值和理论值

正误	词类		合计
	名词	动词	
正确	16 (11)	6 (11)	22
错误	14 (19)	24 (19)	38
合计	30	30	60

根据 χ^2 值表查得自由度为 1，显著性水平 p 为 0.05 和 0.01 时的 χ^2 值分别为 3.84 和 6.63。由于 $\chi^2=7.177>6.63$，则 $p<0.01$，所以应该在 0.01 显著性水平上拒绝虚无假设而接受备择假设。因此结果表明，患者 HW 命名动词图形和名词图形的正确和错误项目的分布模式上是有差异的，她命名名词的正确率显著高于命名动词的正确率。

（二）*t* 检验

t 检验常用于个案（组）—对照组比较研究中。当两组被试的人数相似，并且每组人数大于 10，这时假设被试行为数据的分布是正态的或接近正态的，可以使用 *t* 检验比较两组之间的差异。

在前面提到的汉语阅读障碍儿童的研究中，吴思娜等人（2004）选择了两组儿童：15 名阅读障碍儿童和 14 名对照组儿童。对两组儿童进行了一系列的测试，其中包括语音测验、语素测验、汉字同音判断、汉字语义相关判断等，结果见表 15-13。

表 15-13　两组被试各测验的描述统计及显著性检验结果

	障碍组	对照组	*t* 值
语音测验（正确率）	0.83 (0.1)	0.92 (0.06)	−2.52*
语素测验（正确率）	0.46 (0.14)	0.77 (0.05)	−7.93***
同音判断“是”反应（ms）	990 (284)	728 (117)	3.1**
“否”反应（ms）	1 106 (402)	861 (153)	2.07*
错误率	0.1 (0.09)	0.1 (0.08)	n. s.
语义相关判断（ms）	1 377 (380)	1 043 (271)	2.89**
错误率	0.26 (0.15)	0.12 (0.08)	2.86**

注：n. s. 表示 $p>0.05$，* 表示 $p<0.05$，** 表示 $p<0.01$，*** 表示 $p<0.001$。

描述统计表明，两组儿童在完成各项测验任务上都是有差异的。但是，两组在哪些测验任务上存在的差异是显著的？研究者对障碍组和对照组在各个测验上的平均数之间的差异进一步进行 t 检验发现，在语音意识测验上，障碍组儿童的正确率显著低于对照组儿童（$t_{(27)}=2.52$，$p<0.05$）；在语素意识测验上，障碍组儿童的正确率显著低于对照组儿童（$t_{(27)}=-7.94$，$p<0.001$）。障碍组儿童的语音通达的“是”和“否”反应时（$t_{(27)}=3.1$，$p<0.01$；$t_{(27)}=2.07$，$p<0.05$）、语义相关判断的反应时（$t_{(27)}=2.89$，$p<0.01$）均显著慢于对照组儿童，语义相关判断的错误率显著高于对照组儿童（$t_{(27)}=2.86$，$p<0.01$）。结果表明，阅读障碍儿童的语素意识、语音意识显著落后于正常儿童，在词典通达速度上显著慢于正常儿童，更容易犯错误。

第三节　使用 SPSS 统计数据的方法

本节将结合上一节中的实验数据分析，说明如何应用 SPSS 软件系统进行卡方检验和 t 检验。

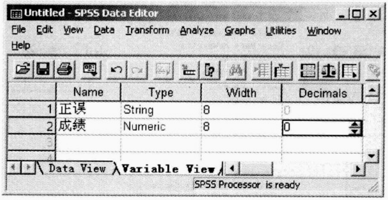
一、卡方检验

（一）拟合度检验

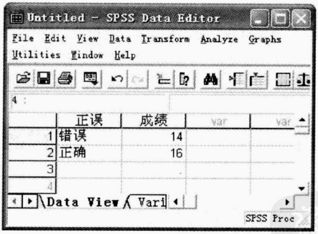
我们仍以患者 HW 的数据为例。他在命名名词图形时，正确和错误项目数分别为 16 个和 14 个。现在我们想检验他的正确与错误项目数的比例是否符合 9∶1。

1. 定义变量、输入变量

首先，进入 SPSS 系统的数据编辑器，并打开 Variable View 窗口。然后定义两个变量“正误”和“成绩”，在数据类型 Type 一栏中，把正误（分类变量）和成绩（数值变量）分别定义为字符型数据（即 String）和数值型数据（即 Numeric），把小数位数（即 Decimals）都设为 0 个，如下图。

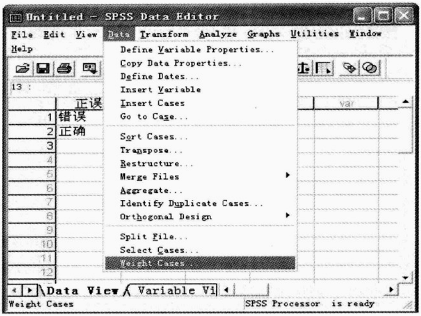


然后打开 Date View 窗口，在“正误”一栏中输入正确和错误及其对应的成绩。值得注意的是，这时的数据要按照数值由小到大的顺序依次输入。

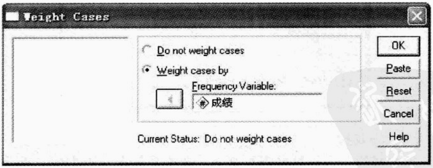


2. 数据加权

对“成绩”一栏数据进行加权，做法是依次点击 Date、Weight Cases。

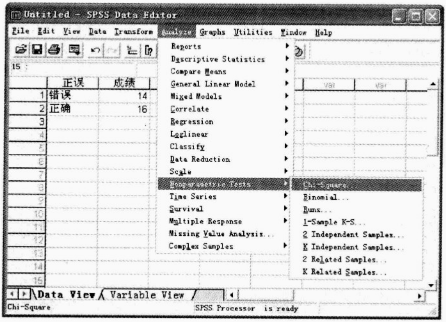


接下来，便会弹出如下对话框。随后点击 Weight case by，将变量“成绩”选入到 Frequency Variable 光带，随后点击 OK 钮返回到 SPSS 的数据窗。需要注意的是，这里加权的只是实测值（成绩），而不是分类变量（正误）。

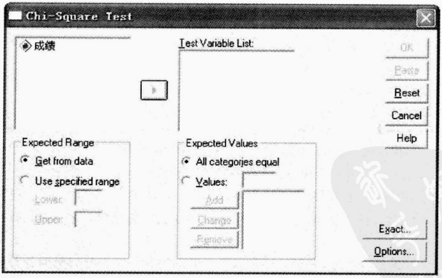


3. 数据检验

依次点击 Analyze、Nonparametric Tests、Chi-Square。

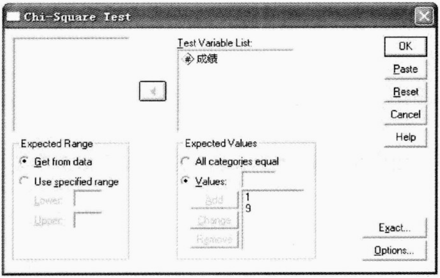


随后会弹出如下对话框。



接下来，把“成绩”选入到 Test Variable List 选项，并在 Expected Values 一栏中选中 Values：依次输入 1 和 9。需要注意的是，Expected

Values 选项是要输入各项目理论值的具体比例。如果各项目的理论值比例是相等的，也就是 1∶1，那么你可以直接选 All categories equal。但如果各项目的期望值比例不相等，就需要选取 Values 选项，并将你的比例数依次输入。特别注意的是，这时输入的比例数的先后顺序要与你的数据表中数据的顺序要相同。如在本例中，正误的选项中第一行错误（14）和第二行正确（16），我们旨在检验它们的比例是否符合 1∶9。所以，我们要先输 1，再输 9，而不能先输 9，再输 1。除此之外，我们还可以在这个对话框中选取一些其他选项，如进入下一级 Options 对话框，能根据自己的意愿选择合适的条件。



随后点击 OK 钮，就会显示如下的统计检验结果（见下页图 15-4）。

图 15-4 中第一个表显示了实测值及其预期值，第二个表显示了卡方检验的结果。可以看出卡方值为 44.815，自由度为 1，p 为 0.000，说明 HW 的命名正确和错误项目数不符合 9∶1 的分布比例，$\chi^2_{(1)}=44.815$，$p<0.001$。

如果我们感兴趣的话，还可以算一下 HW 的正确与错误数的分布是否符合 1∶1 的比例。检验结果发现 HW 的正确与错误数的分布是符合 1∶1的比例的，$\chi^2_{(1)}=0.133$，$p=0.715>0.05$（见下页图 15-5）。

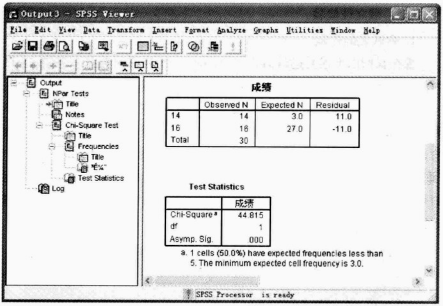


图 15-4　实测值与预期值（1∶9）的拟合度检验输出结果

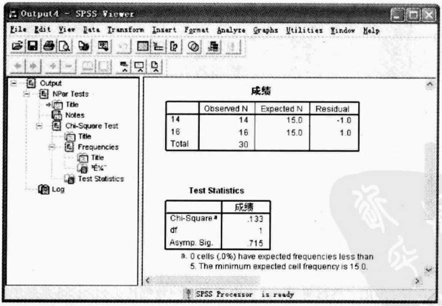


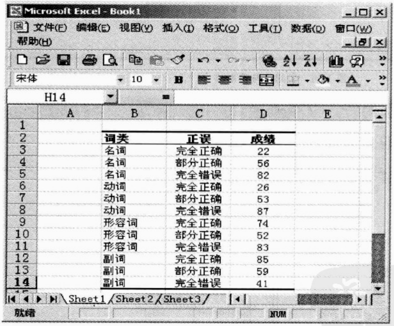
图 15-5　实测值与预期值（1∶1）的拟合度检验输出结果

（二）独立性检验

1. 列联表的检验

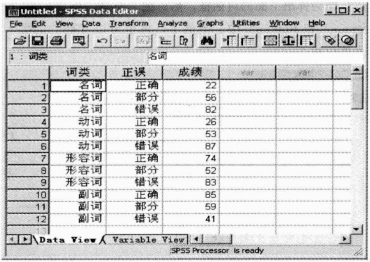
现在我们以上述复述障碍患者 FS 的 4×3 列联表数据（见表 15-11）为例，说明如何用 SPSS 软件系统来进行列联表检验。

(1) 整理数据。如果数据很复杂，我们可以在 Excel 表中整理好后再转入 SPSS 数据表。如果数据比较简单，可以在 SPSS 数据表中直接输入。对于表 15-11 数据来说，在 Excel 表中整理为如下形式。其中，“词类”和“正误”均属于分类变量，而不是连续变量。这里需要注意的是，“成绩”一栏要填写正确或错误的具体数目，而不是项目的总数或比例。



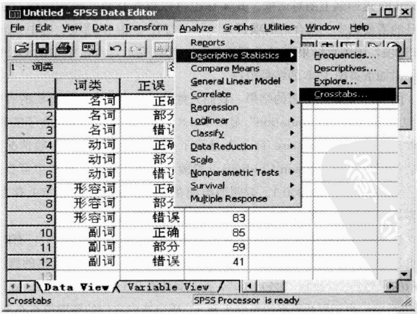
词类	正误	成绩
名词	完全正确	22
名词	部分正确	56
名词	完全错误	82
动词	完全正确	26
动词	部分正确	53
动词	完全错误	87
形容词	完全正确	74
形容词	部分正确	52
形容词	完全错误	83
副词	完全正确	85
副词	部分正确	59
副词	完全错误	41

(2) 定义变量、输入变量。启动 SPSS 系统，在 Variable View 窗口中定义三个变量：词类、正误和成绩，且前两个变量是字符型的，最后一个是数值型的。然后将 Excel 表中整理好的数据拷贝到 SPSS 系统中的 Data View 表中。

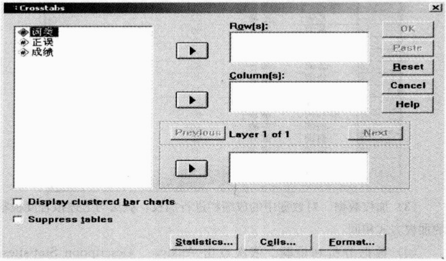


（3）加权数据。对数据中的成绩栏进行加权，做法与上述拟合度检验的加权方式相同。

（4）数据分析和检验。依次点击 Analyze、Description Statistics、Crosstabs。



随后，便会弹出下面的对话框。然后把“词类”和“正误”分别选入到 Column〔s〕和 Row〔s〕中。当然了，你也可以互换“词类”和“正误”的入选位置。

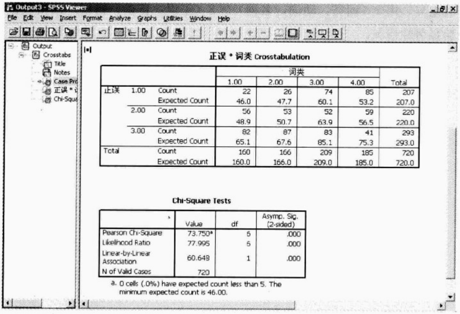


然后，打开二级 Statistics 对话框，选中 Chi-Square 选项，并点击 Continue 返回到上图对话框。其实你也可以打开其他二级对话框，根据自己的需要选中一些合适的项目。比如，打开 Cells 对话框，选中 expected 选项，则表示结果要显示理论值。一切选完后，点击图中的 OK 钮，便弹出如下页图 15-6 的结果窗口。

图 15-6 是数据分析的结果，我们只截选了其中两个比较重要的表格。“词类 * 正误 Crosstabutation”显示了 4×3 的列联表，其中包含了实测值和理论值等。“Chi-Square Tests”显示了卡方检验的结果。一般情况下，我们使用第一行 Pearson Chi-Square 检验结果就可以了。从图中可看出，$\chi^2_{(6)}=73.750$，$p<0.0001$。表明四类词在三种正误类型的分布上存在显著差异，也就是词类对患者的复述成绩有明显影响。

这里需注意的问题是，如果我们的实测值非常小（如 1、3），有时便会出现理论值小于 5 的情形，这时便发现 Chi-Square Tests 表格的脚注 b 显示的不再是“0 cells (0%) ...”，而是具体有几个理论值小于 5。如下页图 15-7 中，由于原始数据中包含 1 和 4，所以造成它们的预期值也小于 5，

脚注 b 为“2cells（50%）...”。在这种情况下，卡方检验的最终结果就不能直接采用皮尔逊检验的值，而是要用对其校正后的值，即表中的 Continuity Correction 一栏的数据，可表示为 $\chi^2_{(1)}=0.873$，$p=0.350>0.05$。简言之，如果有一个或多个预期值小于 5，检验的结果就应该用校正后的值。



正误 * 词类 Crosstabulation

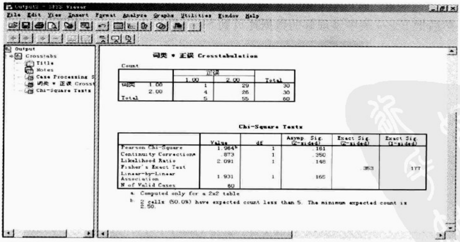
			词类				
			1.00	2.00	3.00	4.00	Total
正误	1.00	Count	22	26	74	85	207
		Expected Count	46.0	47.7	60.1	53.2	207.0
	2.00	Count	56	53	52	59	220
		Expected Count	48.9	50.7	63.9	56.5	220.0
	3.00	Count	82	87	83	41	293
		Expected Count	65.1	67.6	85.1	75.3	293.0
Total		Count	160	166	209	185	720
		Expected Count	160.0	166.0	209.0	185.0	720.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	73.750[a]	6	.000
Likelihood Ratio	77.995	6	.000
Linear-by-Linear Association	60.648	1	.000
N of Valid Cases	720		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 46.00.

图 15-6　实测值和理论值的列联表检验结果



词类 * 正误 Crosstabulation

Count

		正误		
		1.00	2.00	Total
词类	1.00	1	29	30
	2.00	4	26	30
Total		5	55	60

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.964[b]	1	.161		
Continuity Correction[a]	.873	1	.350		
Likelihood Ratio	2.091	1	.148		
Fisher's Exact Test				.353	.177
Linear-by-Linear Association	1.931	1	.165		
N of Valid Cases	60				

a. Computed only for a 2x2 table

b. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 2.50.

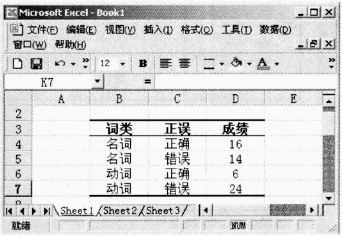
图 15-7　卡方检验结果的校正

2. 四格表的检验

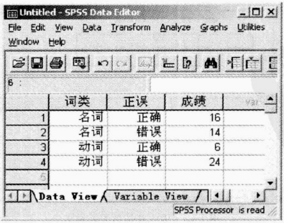
在 SPSS 软件系统中，四格表的检验步骤与列联表的检验步骤基本相同。

现在，我们再以 HW 患者为例来简单地说明如何进行四格表的检验。我们想看一看 HW 在图形命名中对不同词类词的命名的正确和错误是否有差异，用 2（词类：动词，名词）×2（正误：正确，错误）四格表检验显著性情况。

（1）整理数据。在 Excel 表中，将实测值可整理为如下形式。

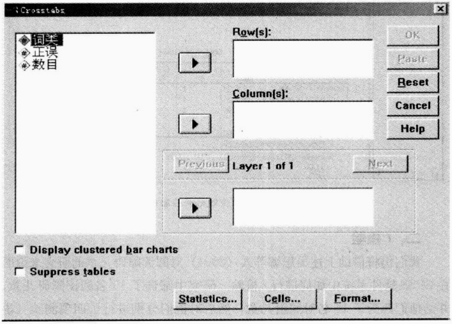


（2）定义变量、输入变量。启动 SPSS 系统，定义三个变量：词类、正误和成绩，而且把前两个变量定义为字符型，最后一个定义为数值型。然后把 Excel 表中整理好的数据拷贝到 SPSS 数据表中，形式如下。

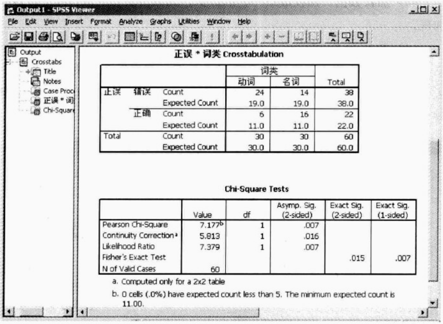


（3）加权数据。加权成绩一栏的数据，具体做法参见拟合度检验数据加权。

（4）数据分析和检验。依次点击 Analyze、Description Statistics、Crosstabs，便会弹出下面的对话框。然后把“正误”和“词类”分别选入到 Row〔s〕 和 Column〔s〕 中。并打开 Statistics 窗口选中 Chi-Square 选项。



然后，点击图中的 OK 钮，检验的结果就会显示出来（见下页图 15-8）。从图中可以看出，$\chi^2=7.177$，$p=0.007<0.01$，表明 HW 对名词的命名正确率显著高于对动词的命名正确率。



正误 * 词类 Crosstabulation

			词类		
			动词	名词	Total
正误	错误	Count	24	14	38
		Expected Count	19.0	19.0	38.0
	正确	Count	6	16	22
		Expected Count	11.0	11.0	22.0
Total		Count	30	30	60
		Expected Count	30.0	30.0	60.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	7.177[b]	1	.007		
Continuity Correction[a]	5.813	1	.016		
Likelihood Ratio	7.379	1	.007		
Fisher's Exact Test				.015	.007
N of Valid Cases	60				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 11.00.

图 15-8　实测值与预期值四格表检验的输出结果

二、*t* 检验

我们仍将借助上述吴思娜等人（2004）对阅读障碍儿童的研究来说明在 SPSS 软件系统中如何进行 *t* 检验。研究中筛选了 15 名阅读障碍儿童，并给他们匹配了 14 名对照组儿童，然后对他们分别进行了四项测验（语音测验、语素测验、同音判断、语义相关判断），测得的结果见表 15-14。

表 15-14　每名儿童在四项测验中测试结果的原始数据

障碍组						对照组					
被试	语素意识（正确率）	语音意识（正确率）	同音判断（毫秒）“Y”反应	同音判断（毫秒）“N”反应	语义相关判断(毫秒)	被试	语素意识（正确率）	语音意识（正确率）	同音判断（毫秒）“Y”反应	同音判断（毫秒）“N”反应	语义相关判断(毫秒)
1	0.28	0.89	1 736	2 404	1 187	1	0.7	0.97	762	761	1 119
2	0.47	0.69	582	740	908	2	0.81	0.97	540	631	924
3	0.58	0.78	798	795	936	3	0.78	0.81	698	978	914

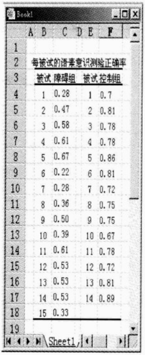
续表

障碍组						对照组					
被试	语素意识（正确率）	语音意识（正确率）	同音判断（毫秒）“Y”反应	同音判断（毫秒）“N”反应	语义相关判断（毫秒）	被试	语素意识（正确率）	语音意识（正确率）	同音判断（毫秒）“Y”反应	同音判断（毫秒）“N”反应	语义相关判断（毫秒）
4	0.61	0.83	1 157	1 048	1 572	4	0.78	0.94	601	678	731
5	0.67	0.97	1 326	1 322	1 653	5	0.86	0.86	856	1 177	1 087
6	0.22	0.86	981	1 151	1 712	6	0.81	0.94	810	967	857
7	0.28	0.69	589	612	716	7	0.72	1.00	841	935	1 064
8	0.36	0.92	917	965	1 151	8	0.75	0.83	740	811	989
9	0.50	0.86	930	1 094	1 449	9	0.75	0.86	805	881	655
10	0.39	0.78	1 131	1 181	1 535	10	0.67	0.94	580	761	1 094
11	0.61	0.81	791	877	1 427	11	0.78	0.94	567	650	1 260
12	0.53	0.67	846	922	1 380	12	0.72	0.86	737	955	795
13	0.53	0.97	1 218	1 376	1 487	13	0.81	1.00	913	937	1 535
14	0.53	0.78	966	1 163	2 028	14	0.89	0.89	752	934	1 572
15	0.33	1.00	886	947	1 510						

如果我们现在想知道两组被试在语素意识测验上差异是否明显，那么就需要对以上实测值进行统计检验。在这种情况下，我们便可以采用 t 检验的方法，其操作步骤如下。

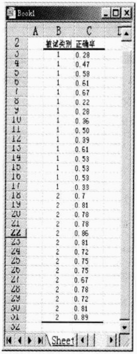
1. 整理数据

在 Excel 表中把原始数据可以整理为下面图 A 的形式，然后把它进一步整理为两列，一列为被试的正确率，另一列为两组被试的分类编号。我们可以把障碍组和对照组分别以“1”和“2”的编码代替，这样就可以整理为图 B 的形式。

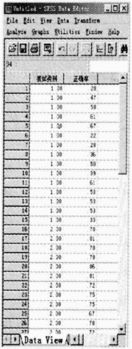


被试	障碍组	被试	控制组
1	0.28	1	0.7
2	0.47	2	0.81
3	0.58	3	0.78
4	0.61	4	0.78
5	0.67	5	0.86
6	0.22	6	0.81
7	0.28	7	0.72
8	0.36	8	0.75
9	0.50	9	0.75
10	0.39	10	0.67
11	0.61	11	0.78
12	0.53	12	0.72
13	0.53	13	0.81
14	0.53	14	0.89
15	0.33		

A



B



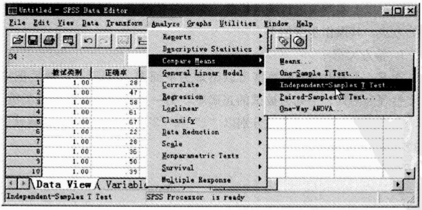
C

2. 定义变量、输入变量

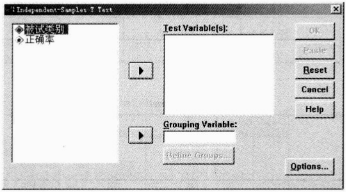
启动 SPSS 软件系统，定义两个变量：被试类别，正确率。然后将 Excel 表中的数据拷贝到 SPSS 系统中的数据窗中，如上面的图 C 的形式。

3. 分析数据

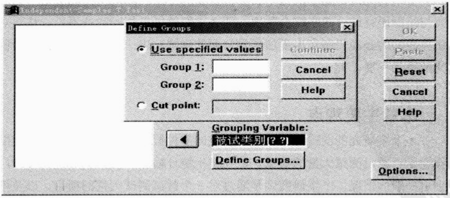
依次点击 Analyze、Compare Means、Independent-Sample T Test。



随后，便会弹出如下图的对话框。然后把“被试类别”和“正确率”分别选入到 Grouping Variable 和 Test Variable〔s〕中。这里需要注意的是，“被试类别”和“正确率”的人选位置不能互换。



接下来，选中下图中的被试类别，并点击 Define Groups 按钮，便出现如下对话框。随之，选中 Use specified values 选项，在 Group1 和 Group2 中分别填入“1”和“2”。这里需要说明的是，所填的“1”和“2”是第一组和第二组的分类变量值，这要与数据窗口的分组变量值一致。另外，如果我们选择了 Cut point，则应该在后面的矩形框中输入一个分组变量的值，以该值为分界线，把大于该值和小于该值各归为一组。就本文数据来说，如果我们选择了 Cut point，则选择 1～2 之间的任何一个数值均可以（如 1.83），它也是把等于 1 和 2 的数据各归为一组。



分别点击上图的 Continue 钮和 OK 钮，就会输出下面的结果（见表 15-15）。

表 15-15 t 检验的输出结果

Group Statistics

	被试类别	N	Mean	Std. Deviation	Std. Error Mean
正确率	1.00	15	.4593	.1399	3.613E-02
	2.00	14	.7736	6.084E-02	1.626E-02

Independent Samples Test

		Levene's Test for quality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
正确率	Equal variances assu	13.251	.001	-7.741	27	.000	-.3142	4.060E-02	-.3975	-.2309
	Equal variances not assumed			-7.932	19.390	.000	-.3142	3.962E-02	-.3970	-.2314

从表中可看出，$t=-7.932$，$p<0.001$，表明障碍组与对照组在语素意识任务差异极显著，具体表现为障碍组的正确率（46%）明显低于对照组的正确率（77%）。

根据同样的方法，我们可以考察两组被试在其他几个测验（语音测验、同音判断和语义相关判断）上差异是否明显，并将检验的结果与表 15-13 相对照。从表 15-13 中可以看出，障碍组在几个测验上的成绩均显著差于对照组。

个案研究是心理学研究中的一种重要的方法。如果我们能够把个案研究与组群研究结合起来综合利用，那么它们就可以相互弥补不足，从而对机体内部的心理机制作出更加科学全面的理解，进一步推动心理学实验研究的蓬勃发展。

本章主要观点

• 个案研究的前提假设与组群研究的前提假设是类似的。组群研究的假设是，通过测试大量的被试，以及推论统计检验，使研究结论能够推广到更一般的情况。个案研究则是通过对一个被试测试大量的项目，以及推论统计检验，使研究结论能够推广到更一般的情况。

• 个案研究是组群研究的补充，它不仅对描述一些稀少事件、个体差异大的心理现象非常重要，而且也对建立理论模型发挥了积极作用。

• 个案—对照组研究方法的基本方法是，为个案设立一个在年龄、性别、受教育程度等无关因素方面匹配的正常对照组。研究者可以通过个案和对照组被试在完成各种任务上的比较，揭示个案的异常的心理机制。

• 分离和相关的方法是认知神经心理学中运用比较普遍的方法。两种方法主要借助患者自身在完成不同任务之间，或通过患者与患者完成任务的相互比较，来推知患者某种认知功能的损伤或保留情况。分离的方法有助于区分两个相对独立的认知加工环节；相关的方法有助于找出两个功能间共同的认知环节。

• 个案研究的数据有明显不同于一般组群研究的数据的特点。已经发展出了多种检验个案研究结果的方法，其中最常用的是卡方检验和 t 检验。卡方检验是处理离散数据的一种检验方法，主要考察实测值和理论值的吻合程度，可分为拟合度检验和独立性检验。t 检验常用于个案（组）—对照组比较研究中。

思考题

1. 什么是个案研究？组群研究和个案研究的特点和适合于研究的问题有什么异同？
2. 个案研究的基本思想和假设是什么？
3. 什么是个案研究的分离方法？什么是单分离？什么是双分离？
4. 什么是个案研究的相关方法？
5. 什么是个案—对照组方法和障碍组—对照组方法？
6. 举例说明卡方检验在个案研究中的运用。
7. 举例说明 t 检验在个案研究中的运用。

参考文献

程书肖、李仲来编著：《教育统计方法》，辽宁大学出版社，1988。

冯伯麟著：《教育统计学》，人民教育出版社，2005。

郝德元编著：《教育与心理统计》，教育科学出版社，1982。

刘友谊：《获得年龄及其汉语视觉词汇加工中的作用机制》，北京师范大学博士论文，2006。

栾辉、舒华、黎程正家、林薇：《汉语发展性深层阅读障碍的个案研究》，载《心理学报》，2003（4），338～343页。

彭聃龄主编：《普通心理学》，北京师范大学出版社，2001。

舒华编著：《心理与教育研究中的多因素实验设计》，北京师范大学出版社，1994。

舒华、储齐人、孙燕、李翔：《移动窗口条件下阅读过程中字词识别特点的研究》，载《心理科学》，1996（2），79～83页。

舒华、张厚粲、程元善：《235个图形的命名一致性、熟悉性、表象性和视觉复杂性评定》，载《心理学报》，1989（4），389～396页。

舒华、张学民、韩在柱著：《实验心理学的理论、方法与技术》，人民教育出版社，2006。

孙汉银：《中文易懂性公式》，北京师范大学硕士论文，1985。

孙宏林、黄建平、孙德金、李德均、邢红兵：《“现代汉语研究语料库系统”概述》，见胡明扬主编《第五届世界汉语教学讨论会论文选》，北京大学出版社，1997。

武宁宁、舒华：《无语境条件下汉语词类歧义词的意义激活》，载《心理学报》，2001（4），305～311页。

吴思娜、舒华、王彧：《汉语发展性阅读障碍的异质性研究》，载《心理发展与教育》，2004（3），46～50页。

张厚粲、徐建平著：《现代心理与教育统计学》，北京师范大学出版社，2004。

张雷、雷雳、郭伯良著：《多层线性模型应用》，教育科学出版社，2002。

张亚旭、周晓林、闵保全、贾建平：《范畴特异性损伤与人脑中一般知识的组织》，载《心理科学》，2003（4），698～700页。

Brace, N., Kemp, R. & Snelgar, R. (2000), *SPSS for psychologists: A*

guide to data analysis using SPSS for Windows. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bridgman, P. W. (1927), *The logic of modern physics*. New York: Macmillan.

Caramazza, A., Chialant, D., Capasso, R. & Micelli, G. (2000). Separable processing of consonants and vowels. *Nature*, 403, 428-430.

Caramazza, A. & Hillis, A. E. (1991), Lexical organization of nouns and verbs in the brain. *Nature*, 349, 788-790.

Caramazza, A. &Shelton, J. R. (1998), Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience*, 10, 1-34.

Chen, L., Zhang, S. & Srinivasan, M. V. (2003), Global perception in small brains: Topological pattern recognition in honey bees. *Proceedings of the National Academy of Sciences of the United States of America*, 100, 6884-6889.

Eron, L. D., Huesman, L. R., Lefkowitz, M. M. & Walder, L. O. (1972), Does television violence cause aggression? *American Psychologist*, 27, 253-263.

Emde, R. N., Plomin, R., Robinson, J., et al. (1992), Temperament, emotion, and cognition at fourteen months: The MacArthur longitudinal twin study. *Child Development*, 63, 1437-1455.

Forster, K. I. & Forster, J. C. (2003), DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods*, *Instruments & Computers*, 35 (1), 116-124.

Gehring, W. J. & Willoughby, A. R. (2002), The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, 295, 2279-2282.

Gilligan, C. (1982), *In a different voice: Psychological theory and women's development*. Cambrige, MA: Harvard University Press.

Goodwin, C. J. (1995), *Research in psychology: Methods and design*. New York: John Wiley & Sons, Inc.

Goodwin, C. J. (1998), *Research in psychology: Methods and design*. New York: John Wiley & Sons, Inc.

Hays, W. L. (1988), *Statistics*. Holt, Rinehart and Winston, Inc.

Hillis, A. E. & Caramazza, A. (1995), Representation of grammatical categories of words in the brain. *Journal of Cognitive Neuroscience*, 7, 396-407.

Keenan, J. P., Nelson, A., O'Connor, M., et al. (2001), Self-recognition

and the right hemisphere. *Nature*, 409, 305.

Kirk, R. E. (1982), *Experimental design: Procedures for the behavioral sciences*. Wadsworth Inc.

Klein, R. M. (1988), Inhibitory tagging system facilitates visual search. *Nature*, 334, 430-431.

Kohlberg, L. (1964), Development of moral character and moral behavior. In L. W. Hoffman & M. L. Hoffman (Eds.), *Review of child development research* (Vol. 1). New York: Sage Publications.

Koutstaal, W. & Rosenthal, R. (1994), Contrast analysis in behavioral research. In J. Brzezinski (Ed.), *Probability in theory-building: Experimental and non-experimental models of scientific research in behavioral sciences*. Amsterdam: Rodopi.

Lenneberg E. (1962), Understanding language without ability to speak: A case report. *Journal of Abnormal and Social Psychology*, 65, 419-425.

Levine, G. & Parkinson, S. (1994), *Experimental methods in psychology*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Luck, S. J., Woodman, G. F. & Vogel, E. K. (2000), Event-related potential studies of attention. *Trends in Cognitive Sciences*, 4, 432-440.

May, C. P. & Hasher, L. (1998), Synchrony effects in inhibitory control over thought and action. *Journal of Experimental Psychology: Human Perception and Performance*, 24 (2), 363-379.

May, C. P., Hasher, L., & Stoltzfus, E. R. (1993), Optimal time of day and the magnitude of age differences in memory. *Psychological Science*, 4, 326-330.

Maylor, E. (1985), Facilitatory and inhibitory components of orienting in visual space. In M. I. Posner & O. S. M. Marin (Eds.), *Attention and performance XI*. Hillsdale, NJ: Erlbaum.

Miller, G. (2002), The good, the bad, and the anterior cingulate. *Science*, 295, 2193-2194.

Morris, R. G., Garrud, P., Rawlins, J. N. & O'Keefe, J. (1982), Place navigation impaired in rats with hippocampal lesions. *Nature*, 297, 681-683.

Petrinovich, L. & Hardyck, C. D. (1969). Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. *Psychological Bulletin*, 71, 43-54.

Posner, M. I. & Cohen, Y. (1984), Components of visual orienting. In H. Bou-

ma & D. Bouwhuis (Eds.), *Attention and performance X*. London: Erlbaum.

Posner, M. I. & Raichle, M. E. (1994), *Images of mind*. New York: Scientific American Library.

Poulton, E. C. (1982), Influential companions: Effects of one strategy on another in the within-subjects designs of cognitive psychology. *Psychological Bulletin*, 91, 673-690.

Rapp, B. & Caramazza, A. (2002), Selective difficulties in spoken nouns and written verbs: A single case study. *Journal of Neurolinguistics*, 15, 373-402.

Rosenthal, R. & Jacobson, L. (1968), *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston.

Rosenthal, R. & Rosnow, R. L. (1991), *Essentials of behavioral research: Methods and data analysis*. McGraw-Hill, Inc.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on psychology's view of human nature. *Journal of Personality and Social Psychology*, 51, 515-530.

Shu, H., Anderson, R. & Wu, N. (2000), Phonetic awareness: Knowledge on orthography-phonology relationship in character acquisition of Chinese children. *Journal of Educational Psychology*, 92 (1), 56-62.

Snodgrass, J. G. & Vanderwart, M. (1980), A standardized set of 260 pictures: Norms of name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 174-215.

Stroop, J. R. (1935), Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643-661.

Su, H, et al. (2003), The Great Wall of China: A physical barrier to gene flow? *Heredity*, 90, 212- 219.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. & Sedivy, J. C. (1995), Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632-1634.

Tang, S. & Guo, A. (2001), Choice behavior of Drosophila facing contradictory visual cues. *Science*, 294, 1543-1547.

Tremblay, R. E., Masse, B., Perron, D., LeBlanc, M., Schwartzman, A. E. & Ledingham, J. E. (1992), Early disruptive behavior, poor school achievement, delinquent behavior, and delinquent personality: Longitudinal analysis. *Journal*

of Consulting and Clinical Psychology, 60, 64-72.

Tversky, A. & Kahneman, D. (1974), Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.

Underwood, B. J. & Shaughnessy, J. J. (1975), *Experimentation in psychology*. New York: Wiley.

Warrington, E. K. & Shallice, T. (1984), Category specific semantic impairments. *Brain*, 107, 829-853.

Watson, J. B. & Rayner, R. (1920), Conditional emotional reactions. *Journal of Experimental Psychology*, 3, 1-14.

Yerkes, R. M. & Dodson, J. D. (1908), The relation of stimulus to rapidity of habit-formation. *Journal of Comparative and Neurological Psychology*, 18, 459-482.

PDG